University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 284

Inference in Additively Separable Models with a High-Dimensional Set of Conditioning Variables

Damian Kozbur

First version: September 2013
This version: April 2018

Inference in Additively Separable Models with a High-Dimensional Set of Conditioning Variables

Damian Kozbur

University of Zürich
Department of Economics
Schönberggasse 1, 8001 Zürich
email: damian.kozbur@econ.uzh.ch

Abstract. This paper studies nonparametric series estimation and inference for the effect of a single variable of interest x on an outcome y in the presence of potentially high-dimensional conditioning variables z. The context is an additively separable model $\mathrm{E}[y|x,z] = g_0(x) + h_0(z)$. The model is high-dimensional in the sense that the series of approximating functions for $h_0(z)$ can have more terms than the sample size, thereby allowing z to have potentially very many measured characteristics. The model is required to be approximately sparse: $h_0(z)$ can be approximated using only a small subset of series terms whose identities are unknown. This paper proposes an estimation and inference method for $g_0(x)$ called *Post-Nonparametric Double Selection* which is a generalization of *Post-Double Selection*. Standard rates of convergence and asymptotic normality for the estimator are shown to hold uniformly over a large class of sparse data generating processes. A simulation study illustrates finite sample estimation properties of the proposed estimator and coverage properties of the corresponding confidence intervals. Finally, an empirical application estimating convergence in GDP in a country-level cross-section demonstrates the practical implementation of the proposed method.

Key Words: additive nonparametric models, high-dimensional sparse regression, inference under imperfect model selection. *JEL Codes*: C1.

1. Introduction

Nonparametric estimation in econometrics and statistics is useful in applications where theory does not provide functional forms for relations between relevant observed variables. In many problems, primary quantities of interest can be computed from the conditional expectation function of an outcome variable y given a regressor of interest x,

$$\mathrm{E}[y|x] = f_0(x).$$

Date: First version: September 2013. This version is of April 16, 2018.
Correspondence: Schönberggasse 1, 8001 Zürich, Department of Economics, University of Zürich, damian.kozbur@econ.uzh.ch.

I thank Christian Hansen, Tim Conley, Matt Taddy, Azeem Shaikh, Dan Nguyen, Dan Zou, Emily Oster, Martin Schonger, Eric Floyd, Kelly Reeve, and seminar participants at University of Western Ontario, University of Pennsylvania, Rutgers University, Monash University, and the Center for Law and Economics at ETH Zurich for helpful comments. I gratefully acknowledge financial support from the ETH Postdoctoral Fellowship.

In this case, nonparametric estimation is a flexible means for estimating unknown f_0 from data under minimal assumptions.

In most econometric models, however, it is also important to take into account conditioning information, given through variables z. Failing to properly control for such variables z will lead to incorrect estimates of the effects of x on y. When such conditioning information is important to the problem, it is necessary to replace the simple objective of learning the conditional expectation function f_0 with the new objective of learning a family of conditional expectation functions

$$\mathrm{E}[y|x,z] = f_{0,z}(x)$$

indexed by z.

This paper studies series estimation[1] and inference of $f_{0,z}$ in a particular case characterized by the following two main features.

1. $f_{0,z}$ is additively separable in x and z, meaning that
$$f_{0,z}(x) = g_0(x) + h_0(z)$$
for some functions g_0 and h_0.
2. The conditioning variables z are observable and high-dimensional.

Additively separable models are convenient in many economic problems because any *ceteris paribus* effect of changing x to x' is completely described by g_0. In addition, a major statistical advantage in restricting to additively separable models is that the individual components g_0, h_0 can be estimated at faster rates than a joint estimation of the family $f_{0,z}$.[2] Therefore, imposing additive separability in contexts where such an assuption is justified is very helpful.

The motivation for studying a high-dimensional framework for z is to allow researchers substantial flexibility in modeling conditioning information when the primary object of interest is g_0. This framework allows analysis of particularly rich or *big* datasets with a large number of conditioning variables.[3] In this paper, high-dimensionality of z is formally defined by the total number of terms in a series expansion of $h_0(z)$. This will allow many possibilities on the types of variables z and functions h_0 covered. For example, z can be high-dimensional itself, while h_0 is approximately linear in the sense that

$$h_0(z) = z_1'\beta_{h_0,1} + ... + z_L'\beta_{h_0,L} + o(1)$$

with $L > n$ and $\beta_{h_0,j}$ denoting the jth component of the vector β_{h_0} and the asymptotic $o(1)$ valid for $L \to \infty$. More generally, z itself can also have moderate

[1]Series estimation of nonparametric regression problems involves least squares estimation performed on a series expansion of the regressor variables. Series estimation is described more fully in Section 2.

[2]Results on faster rates for separable models exist for both kernel methods (marginal integration and back-fitting methods) and series based estimators. For a general review of these issues, see for example the textbook [41]. Additional discussion on the literature on additively separable models is provided later in the introduction.

[3]In many cases, larger set of covariates can lend additional credibility to conditional exogeneity assumptions. See the discussion in [15].

dimension, but any sufficiently expressive series expansion of h_0 must have many terms as a simple consequence of the curse of dimensionality.

A basic mechanical outline for the estimation and inference strategy presented in this paper proceeds in the following steps.

1. Consider approximating dictionaries (equivalently series expansions) with K terms, given by $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))$. Linear combinations of $p^K(x)$ are used for approximating $g_0(x)$. In addition, consider approximating dictionaries with L terms, $q^L(z) = (q_{1L}(z), ..., q_{LL}(z))$, for approximating $h_0(z)$. Possibly $L > n$.
2. Reduce the number of series terms for h_0 in a way which continues to allow robust inference. This requires multiple model selection steps.
3. Proceed with traditional series estimation and inference techniques on the reduced dictionaries.

Strategies of this form are commonly referred to as *post-model selection inference* strategies.

The primary targets of inference considered in this paper are real-valued functionals, $g \mapsto a(g) \in \mathbb{R}$. Specifically, let

$$\theta_0 = a(g_0).$$

Leading examples of such functionals include the average derivative $a(g) = \mathrm{E}[g'(x)]$ or the difference of $a(g) = g(x_0^2) - g(x_0^1)$ for two distinct x_0^1, x_0^2 of interest.

The main contribution of this paper is the construction of confidence sets that cover θ_0 to some pre-specified confidence level. Moreover, the construction is valid uniformly over a large class of data generating processes which allow z to be high-dimensional.

Current high-dimensional estimation techniques provide researchers with useful tools for dimension reduction and dealing with datasets where the number of parameters exceeds the sample size.[4] Most high-dimensional techniques require additional structure to be imposed on the problem at hand in order to ensure good performance. One common structure for which reliable high-dimensional techniques exist is sparsity. Sparsity means that the number of nonzero parameters is small relative to the sample size. In this setting, common techniques include ℓ_1-regularization techniques like Lasso and Post-Lasso[5]. Other techniques include the Dantzig selector, Scad, and Forward Stepwise regression.

The literature on nonparametric estimation of additively separable models is well developed. As mentioned above, additively separable models are useful since they

[4]Statistical models which are extremely flexible, and thus overparameterized, are likely to overfit the data, leading to poor inference and out of sample performance. Therefore, when many covariates are present, regularization is necessary.

[5]The Lasso is a shrinkage procedure which estimates regression coefficients by minimizing a quadratic loss function plus an ℓ_1 penalty for the size of the coefficient. The nature of the penalty gives Lasso favorable property that many parameter values are set identically to zero and thus Lasso can also be used as a model selection technique. Post-Lasso fits an ordinary least squares regression on variables with non-identically-zero estimated Lasso coefficients. For theoretical and simulation results about the performance of these two methods, see [29] [52], [32] [23] [4], [5], [17], [21], [20] [22], [23], [33], [38], [39], [42], [43], [44], [47], [52], [53], [55], [57], [9], [18], [9], among many more.

impose an intuitive restriction on the class of models considered, and as a result provide higher quality estimates. Early study of additively separable models was initiated in [19] and [31], who describe backfitting techniques. [25] propose marginal integration methods in the kernel context. [50] and [56] consider estimation of derivatives in components of additive models. [26] develop local partitioned regression which can be applied more generally than the additive model. In terms of series-based estimation, series estimators are particularly easy to use for estimating additively separable models since series terms can be allocated to respective model components. General large sample properties of series estimators have been derived by [51], [27], [2], [28], [3] [45] [14] and many other references. Relative to kernel estimation, series estimators are simpler to implement, but often require stronger support conditions. Many additional references for both kernel and series based estimation can be found in the reference text [41]. Finally, [34] consider estimation of additively separable models in a setting where there are high-dimensional additive components. The authors propose and analyze a series estimation approach with a Group-Lasso penalty to penalize different additive components. This paper therefore studies a very similar setting to the one in [34], but constructs a valid procedure for forming confidence intervals rather than focusing on estimation error.

The main challenge in statistical inference or construction of confidence intervals after model selection is in attaining robustness to model selection errors. When coefficients are small relative to the sample size (ie statistically indistinguishable from zero), model selection mistakes are unavoidable.[6] Such model selection mistakes can lead to distorted statistical inference in much the same way that pretesting procedures lead to distorted inference. This intuition is formally developed in [46] and [40]. Nevertheless, given the practical value of dimension reduction, and the increasing prevalence of high-dimensional datasets, studying robust post-model selection inference techniques and *post-regularization* inference techniques is an active area of current research. Offering solutions to this problem is the focus of a number of recent papers; see, for example, [11], [8], [58], [12], [15], [54], [36], [10], and [16].[7]

This paper proposes a procedure called *Post-Nonparametric Double Selection* for the additively separable model. The proposed procedure is a generalization of the approach in [15] (named *Post-Double-Selection*). [15] gives robust statistical inference for the slope parameter α_0 of a treatment variable x with high-dimensional control variables z in the context of a partially linear model $\mathrm{E}[y|x,z] = \alpha_0 x + h_0(z)$.[8] The Post-Double Selection method selects elements of z in two steps. Step 1 selects the terms in an expansion of z that are most useful for predicting x. Step 2 selects terms in an expansion of z most useful for predicting y. A consequence of the particular construction using two selection steps is that terms excluded by model selection mistakes twice necessarily have a negligible effect on subsequent statistical inference.[9] Post-Nonparametric Double Selection replaces step 1 of Post-Double

[6]Under some restrictive conditions, for example beta-min conditions which constrain nonzero coefficients to have large magnitudes, perfect model selection can be attained.

[7]Citations are ordered by date of first appearance on arXiv.

[8]Several authors have addressed the task of assessing uncertainties or estimation error of model parameter estimates in a wide variety of models with high dimensional regressors (see, for example, [11], [8], [58], [12], [15], [54], [36], and [10]).

[9]The use of two model selection steps is motivated partially by the intuition that two necessary conditions for omitted variables bias to occur: an omitted variable exists which is (1) correlated with the treatment x, and (2) correlated with the outcome y. Each selection step addresses one

Selection with selecting variables useful for predicting any test function $\varphi(x) \in \Phi_K$ for a sufficiently general class of functions Φ_K.

This paper suggests a simple choice for Φ_K which is based on the linear span of $p^K(x)$. This choice is called $\Phi_{K,\text{Span}}$.[10] Theoretical and simulation results show that the suggested choice has favorable statistical properties uniformly under certain sequences of data generating processes.

Working with a generalization of Post-Double Selection which dissociates the first stage selection from the final estimation is useful for several reasons. One reason is that the direct extension of Post-Double is not invariant to the choice of dictionary $p^K(x)$ and leads natural to the consideration of more general Φ_K. In addition, applying the direct generalization of Post-Double selection may lead to poorer statistical performance than using a larger, more robust Φ_K. A simulation study later in this paper explores these properties. Next, as a theoretical advantage, in some cases a larger Φ_K gives estimates and inference which are valid under weaker rate conditions on K, n, etc. Finally, working dissociating the first stage helps in terms of organizing the arguments in the proofs. In particular, various bounds developed in the proof depend on a notion of density of Φ_K within $\text{LinSpan}(p^K)$.

This paper proves convergence rates and asymptotic normality for Post-Nonparametric Double Selection estimates of $g_0(x)$ and θ_0 respectively. The proofs in the paper proceed by using the techniques in Newey's analysis of series estimators (see [45]) and ideas in Belloni, Chernozhukov, and Hansen's analysis of Post-Double Selection (see [15]), along with careful tracking of a notion of density of the set Φ_K within the linear span of $p^K(x)$. The estimation rates for g_0 obtained in this paper match those of [45]. Next, a simulation study demonstrates finite sample performance of the proposed procedure. Finally, an empirical example estimating the relationship of initial GDP to GDP growth in a cross-section of 90 countries illustrates the use of Post-Nonparametric Double Selection.

2. Series estimation with a reduced dictionary

This section establishes notation, reviews series estimation, and describes series estimation on a *reduced dictionary*. The exposition begins with basic assumptions on the observed data.

Assumption 1 (Data)**.** *The observed data, $\mathcal{D}_n$, is given by n iid copies of random variables $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ indexed by $1 \leqslant i \leqslant n$, so that*

$$\mathcal{D}_n = (y_i, x_i, z_i)_{i=1}^n.$$

Here, y_i are outcome variables, x_i are explanatory variables of interest, and z_i are conditioning variables. In addition, $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^r$ for some integer $r > 0$ and $\mathcal{Z}$ is a general measure space.

---

of the two concerns. In their paper, they prove that under the regularity right conditions, the two described model selection steps can be used to obtain asymptotically normal estimates of α_0 and in turn to construct correctly sized confidence intervals.

[10]Alternative choices are possible and the analysis in the paper covers a general class of choices for Φ_K.

Assumption 2 (Additive Separability)**.** *There is a random variable ε and functions g_0 and h_0 such that the following additive separability*[11] *holds.*

$$y = g_0(x) + h_0(z) + \varepsilon, \quad \mathrm{E}[\varepsilon|x, z] = 0.$$

Traditional series estimation of (g_0, h_0) is carried out by performing least squares regression on series expansions in x and z. Define a dictionary of approximating functions by

$$(p^K(x), q^L(z))$$

where $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))$ and $q^L(z) = (q_{1L}(z), ..., q_{LL}(z))$ are each series of K and L functions such that their linear combinations can approximate $g_0(x)$ and $h_0(z)$. Construct the matrices $P = [p^K(x_1), ..., p^K(x_n)]'$, $Q = [q^L(z_1), ..., q^L(z_n)]'$,$Y = (y_1, ..., y_n)'$, and let $\widehat{\beta}_{y,(K,L)} = ([P\ Q]'[P\ Q])^{-1}[P\ Q]'Y$ be the least squares estimate from Y on $[P\ Q]$. Let $[\widehat{\beta}_{y,(K,L)}]_g$ be the components of $\widehat{\beta}_{y,(K,L)}$ corresponding to p^K. Then $\widehat{g}(x)$ is defined by

$$\widehat{g}(x) = p^K(x)'[\widehat{\beta}_{y,(K,L)}]_g.$$

When $L > n$, quality statistical estimation is only feasible provided dimension reduction or regularization is performed. A dictionary reduction selects new approximating terms,

$$(p^K(x), q^L(z)) \quad \overset{\text{reduction}}{\longrightarrow} \quad (\tilde{p}(x), \tilde{q}(z)),$$

comprised of a subset of the series terms in $(p^K(x), q^L(x))$. In this paper, because the primary objects of interest center around $g_0(x)$, it will be the convention to always take $\tilde{p}(x) = p^K(x)$. The *post-model-selection* estimate of $g_0(x)$ is then defined analogously to the traditional series estimate. Let $\widehat{\beta}_{y,(\tilde{p},\tilde{q})} = ([\tilde{P}\ \tilde{Q}]'[\tilde{P}\ \tilde{Q}])^{-1}[\tilde{P}\ \tilde{Q}]'Y$ where $\tilde{P} = [\tilde{p}(x_1), ..., \tilde{p}(x_n)]' = P$, $\tilde{Q} = [\tilde{q}(z_1), ..., \tilde{q}(z_n)]'$ and as before, let $[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_g$ be the components of $\widehat{\beta}_{y,(\tilde{p},\tilde{q})}$ corresponding to $\tilde{p}$. Then $\widehat{g}$ is defined by

$$\widehat{g}(x) = \tilde{p}(x)'[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_g.$$

Finally, consider a functional[12] $a(g) \in \mathbb{R}$ and as before, set $\theta_0 = a(g_0)$. One sensible estimate for θ_0 is given by

$$\widehat{\theta} = a(\widehat{g}).$$

In order to use $\widehat{\theta}$ for inference on θ_0, an approximate expression for the variance $var(\widehat{\theta})$ is necessary. As is standard, the expression for the variance will be approximated using the delta method. Let $\widehat{A} = \frac{\partial a(p^K(x)'b)}{\partial b}([\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_g)$. Let $\mathcal{M} = \mathrm{Id}_n - \tilde{Q}(\tilde{Q}'\tilde{Q})^{-1}\tilde{Q}'$ be the projection matrix onto the space orthogonal to the

[11]This assumption simply rewrites the equation stated in the introduction in terms of a *residual* ε. To ensure uniqueness of g_0, a further normalization is required. A common normalization in the series context is $g_0(0) = 0$, which is sufficient for most common assumptions on h_0.

[12]Restriction to one dimensional functionals is for simplicity.

span of $\tilde{Q}$.[13] Finally, let $\widehat{\mathcal{E}} = Y - [\tilde{P}, \tilde{Q}]\widehat{\beta}_{y,(\tilde{p},\tilde{q})}$. Estimate $\widehat{V}$ using the following sandwich form:

$$\widehat{V} = \widehat{A}\widehat{\Omega}^{-1}\widehat{\Sigma}\widehat{\Omega}^{-1}\widehat{A}$$
$$\widehat{\Omega} = n^{-1}\tilde{P}'\mathcal{M}\tilde{P}$$
$$\widehat{\Sigma} = n^{-1}\tilde{P}'\mathcal{M}\text{diag}(\widehat{\mathcal{E}})^2\mathcal{M}\tilde{P}.$$

The following sections describe a dictionary reduction technique, along with regularity conditions, which imply that

$$n^{1/2}\widehat{V}^{-1/2}(\theta_0 - \widehat{\theta}) \to_d N(0,1).$$

The practical value of the results is that they formally justify approximate Gaussian inference for θ_0. An immediate corollary of the Gaussian limit is that for any significance level $\gamma \in (0,1)$, with $c_{1-\gamma/2}$ the $(1-\gamma/2)$-quantile of the standard Guassian distribution, it holds that

$$\text{P}(\theta_0 \in [\widehat{\theta} - c_{1-\gamma/2}n^{1/2}\widehat{V}^{-1/2}, \widehat{\theta} - c_{1-\gamma/2}n^{1/2}\widehat{V}^{-1/2}]) \to \gamma.$$

3. Dictionary Reduction by Post-Nonparametric Double Selection

The previous section described estimation using a generic dictionary reduction. This section discusses one class of possibilities for constructing such reductions.

It is important to note that the coverage probabilities of the above confidence sets depend critically on how the dictionary reduction is performed. In particular, naive one-step methods will fail to produce correct inference. Formal results expanding on this point can be found, for instance, in [46], [40]. Heuristically, the reason resulting confidence intervals have poor coverage properties is due to model selection mistakes.

To address this problem, this section proposes a procedure for selecting $\tilde{q}(z)$. The new procedure is a generalization of the methods in [15] who work in the context of the partially linear model $\text{E}[y|x,z] = \alpha_0 x + h_0(z)$. The methods described below rely heavily on Lasso-based model selection. Therefore, a brief description of Lasso is now provided. The following description of Lasso, which uses an overall penalty level as well as term-specific penalty loadings follows [8] who are motivated by allowing for heteroskedasticity.

For any random variable v with observations $(v_1, ..., v_n)$, the *Lasso* estimate v on $q^L(z)$ with *penalty parameter* λ and *loadings* l_j is defined as a solution

$$\widehat{\beta}_{v,L,\text{Lasso}} \in \arg\min \sum_{i=1}^{n}(v_i - q^L(z_i)'b)^2 + \lambda\sum_{j=1}^{L}|l_j b_j|.$$

The corresponding *selected set* $I_{v,L}$ is defined as

$$I_{v,L} = \{j : \widehat{\beta}_{v,L,\text{Lasso},j} \neq 0\}.$$

Finally, the corresponding *Post-Lasso* estimator is defined by

$$\widehat{\beta}_{v,L,\text{Post-Lasso}} \in \arg \min_{b: b_j = 0 \text{ for } j \notin I_{v,L}} \sum_{i=1}^{n}(v_i - q^L(z_i)'b)^2.$$

[13] When the required inverse does not exist, a pseudo-inverse may be used.

Lasso is chosen over other model selection possibilities for several reasons. Foremost, Lasso is a simple, computationally efficient estimation procedure which produces sparse estimates because of its ability to set coefficients identically equal to zero. In particular, $|I_{v,L}|$ will generally be much smaller than n if a suitable penalty level is chosen. The second reason is for the sake of continuity with the previous literature; Lasso was used in [15]. The third reason is for concreteness. There are indeed many alternative estimation or model selection procedures which select a sparse set of terms which in principle can replace the Lasso. It is possible to instead consider general model selection techniques in the course of developing the subsequent theory. However, framing the discussion using Lasso allows explicit calculation of bounds and explicit description of tuning parameters. This is also helpful in terms of practical implementation of the procedures proposed below.

The quality of Lasso estimation is controlled by λ and l_j. As the number of different Lasso estimations increases (ie. with increasingly many different variables v), the penalty parameter must be increased to ensure quality estimation uniformly over all different v. The penalty parameter must also be increased with increasing L. However, higher λ typically leads to more shrinkage bias in Lasso estimation. Therefore, given l_j, λ is usually chosen to be large enough to ensure quality performance, and no larger. See [8] for details.

For the sake of completeness, the Post-Double Selection procedure of [15] is now reproduced for a partially linear model specified by $\mathrm{E}[y|x,z] = \alpha_0 x + h_0(z)$.

Algorithm 1. *Post-Double Selection for the Partially Linear Model.* (Reproduced from [15]).

1. *First Stage Model Selection Step.* Perform Lasso regression x on $q^L(z)$ with penalty λ_{FS} and loadings $l_{\mathrm{FS},j}$. Let I_{FS} be the set of selected terms.

2. *Reduced Form Model Selection Step.* Perform Lasso regression y on $q^L(z)$ with penalty λ_{RF} and loadings $l_{\mathrm{RF},j}$. Let I_{RF} be the set of selected terms.

3. *Post-Model Selection Estimation.* Set $I_{\mathrm{PD}} = I_{\mathrm{FS}} \cup I_{\mathrm{RF}}$ and let $\tilde{q}(z) = [q_{jL}(z)]_{j \in I_{\mathrm{PD}}}$. Estimate α_0 with $\widehat{\alpha}$ based on least squares regression[14] of y on $[x, \tilde{q}(z)]$.

Appendix A contains details about one possible method for choosing $\lambda_{\mathrm{FS}}, \lambda_{\mathrm{RF}}$ as well as $l_{\mathrm{FS},j}$,$l_{\mathrm{RF},j}$. Arguments in [15] show that the choices of tuning parameters given in Appendix A are sufficient to guarantee a centered Gaussian sampling distribution of $\widehat{\alpha}$ for α_0.

The simplest generalization of Post-Double Selection is to expand the first stage selection step into K steps. More precisely, for $k = 1, ..., K$, perform Lasso regression of $p_{kK}(x)$ on $q^L(z)$, and set $I_{\mathrm{FS},k}$ as the selected terms. Then define $I_{\mathrm{FS}} = \cup_{k=1}^{K} I_{\mathrm{FS},k}$ and continue to the reduced form and estimation steps.[15] This approach has a few disadvantages. First, the selected variables can depend on the particular dictionary $p^K(x)$. Ideally, the first stage model selection should be approximately invariant to the choice of $p^K(x)$.

[14] In [15] heteroskedasticity-consistent standard errors are used for inference.

[15] A previous draft of this paper took this approach. Deriving theoretical results for this approach requires stronger sparsity assumptions than required here.

Instead, consider a general class of test functions $\Phi_K = \{\varphi\}$. Concrete classes for test functions are provided below. In the first stage in Post-Nonparametric Double Selection, a Lasso step of $\varphi(x)$ on $q^L(z)$ is performed for each $\varphi \in \Phi_K$.

Algorithm 2. *Post-Nonparametric Double Selection*

1. *First Stage Model Selection Step.* For each $\varphi \in \Phi_K$, perform Lasso regression $\varphi(x)$ on $q^L(z)$ with penalty λ_φ and loadings $l_{\varphi,j}$. Let $I_{\varphi,L}$ be the selected terms. Let $I_{\Phi_K} = \cup_{\varphi\in\Phi_K} I_{\varphi,L}$ be the union set of selected terms.

2. *Reduced Form Model Selection Step.* Perform Lasso regression y on $q^L(z)$ with penalty λ_{RF} and loadings $l_{\text{RF},j}$. Let I_{RF} be the set of selected terms.

3. *Post-Model Selection Estimation.* Set $I_{\Phi_K+\text{RF}} = I_{\Phi_K} \cup I_{\text{RF}}$. Estimate θ_0 using $\widehat{\theta}$ based on the reduced dictionary
$$(\tilde{p}(x), \tilde{q}(z)) = (p^K(x), [q_{jL}(z)]_{j\in I_{\Phi_K+\text{RF}}}).$$

The following are several concrete, feasible options for Φ_K. The first option is named the Span option. This option is suggested for practical use and is the main option in the simulation study as well as in the empirical example that follow.

$$\Phi_{K,\text{ Span}} = \{\varphi(x) \in \text{LinSpan}(p^K(x)) : \text{var}(\varphi(x)) \leqslant 1\}.$$

The theory in the subsequent section is general enough to consider other options for Φ_K which might possibly be preferred in different contexts. Three additional examples are as follows.

$$\Phi_{K,\text{ Graded}} = \{p_{11}(x)\} \cup \{p_{12}(x), p_{22}(x)\} ... \cup \{p_{1K}(x), ..., p_{KK}(x)\}$$
$$\Phi_{K,\text{ Multiple}} = \{p^{(1)}_{1K}(x), ..., p^{(1)}_{KK}(x)\} \cup ... \cup \{p^{(m)}_{1K}(x), ..., p^{(m)}_{KK}(x)\}\}$$
$$\Phi_{K,\text{ Simple}} = \{p_{1K}(x), ..., p_{KK}(x)\}.$$

Appendix A again contains full implementation details for the Span option. This includes one possible method for choosing λ_φ, $l_{\varphi,j}$ as well as $l_{\text{FS},j}$,$l_{\text{RF},j}$ which yield favourable model selection properties. Discussion of the most important details is given in the text below. The analysis in the next section gives conditions under which $\widehat{\theta}$ attains a centered Gaussian limiting distribution.

Choosing Φ_K optimally is an important problem, which is similar to the problem of dictionary selection.[16] The Span option, $\Phi_{K,\text{ Span}}$ is used in the simulation study as well as in the empirical example, since it performed well in initial simulations. Note that the definition of the set $\Phi_{K,\text{ Span}}$ depends on a population quantity $\text{var}(\varphi(x))$ which may be unknown to the researcher. Note however, that the identities of the covariates selected in the Lasso-based procedure described in the appendix are invariant to rescaling of the left-hand side variable. The invariance

[16]The question of which option for Φ_K is optimal is likely application dependent. In order to maintain focused, this question is not considered in detail in this paper but might be of interest for future work.

is a consequence of the method for choosing penalty loadings. Therefore, replacing the condition $\sup_{x\in\mathcal{X}} \|\varphi(x)\|_2 \leqslant 1$ with $\text{var}(\varphi(x)) \leqslant 1$ is possible. The option $\Phi_{K,\text{ Simple}}$ is the direct extension of Post Double Selection as given in [15]. The set $\Phi_{K,\text{ Multiple}}$ corresponds to using multiple dictionaries, indexed $(1),...,(m)$ in the notation above. For example $\Phi_{K,\text{ Multiple}}$ could include the union of B-splines, orthogonal polynomials, and trigonometric polynomials, all in the first stage selection. The $\Phi_{K,\text{ Graded}}$ is appropriate when dictionaries are not nested with respect to K. These include B-splines.

In order to set up a practical choice of penalty levels, the set proposed above, Φ_K is considered as a union[17] :

$$\Phi_{K,\text{Span}} = \Phi_{K1} \cup \Phi_{K2} \cup \Phi_{K3}$$

where

$$\begin{aligned}
&\Phi_{K1} = \{x\} \\
&\Phi_{K2} = \{p_{1K}(x),...,p_{KK}(x)\} \\
&\Phi_{K3} = \{\varphi(x) \in \text{LinSpan}(p^K(x)) : \text{var}(\varphi(x)) \leqslant 1\}.
\end{aligned}$$

The reason then for decomposing $\Phi_{K,\text{Span}}$ in this way is allow the use of different penalty levels on each of the three sets $\Phi_{K1}, \Phi_{K2}, \Phi_{K3}$. In particular, $\lambda_{\Phi_{K1}}$ is the penalty for a single heteroskedastic Lasso as described in [8]. $\lambda_{\Phi_{K2}}$ is a penalty which adjusts for the presence of K different Lasso regressions with $K \to \infty$. The main proposed estimator sets $\lambda_{\Phi_{K3}} = \lambda_{\Phi_{K2}}$. This is less conservative than the penalty level would be following [10] for a continuum of Lasso estimations.[18] As a result, any corresponding Lasso performance bounds do not hold uniformly over Φ_{K3}. Rather the implied bounds hold only uniformly over any pre-specified K element subsets of Φ_{K3}. The high-level model selection assumption below (see Assumption 10) indicates that these bounds are sufficient for the present purpose. In the simulation study, a more conservative (higher) choice for $\lambda_{\Phi_{K3}}$ is also considered. In terms of inferential quality, there is no noticeable difference between the two choices of penalties in the data generating processes considered in the simulation study. As discussed above, penalty levels accounting for a set of different Lassos estimated simultaneously must be higher to ensure quality estimation. This leads to higher shrinkage bias. The above decomposition therefore addresses both concerns about quality estimation and shrinkage bias by allowing smaller penalty levels to be used on subsets of $\Phi_{K,\text{Span}}$. Because the decomposition is into a fixed, finite number of terms (ie. into 3 terms), such an estimation strategy presents no additional theoretical difficulties.

Another practical difficulty with this approach is computational. It is infeasible to estimate a Lasso regression for every φ indexed by a continuum. Therefore, some approximation must be made. The reference [10] gives suggestions for estimating a continuum of Lasso regressions using a grid. This may be computationally expensive if K is even moderately large. An alternative heuristic approach is motivated by the observation that q_{jL} is selected into I_{Φ_K} only when there is $\varphi \in \Phi_K$ such that $j \in I_{\varphi,L}$. In the context of estimating θ_0, only the identity of selected terms is

[17]Some dictionaries $p^K(x)$ may not contain a term $p_{kK}(x) = x$. In this case, $\varphi(x) = x$ can be appended to Φ_K. In addition, after rescaling, $\Phi_{K1} \subseteq \Phi_{K2} \subseteq \Phi_{K3}$ is possible, and so the sets have nonempty intersection. This causes no additional problems.

[18]Note, the normalization that $\|\varphi(x)\|_2 \leqslant 1$ ensures that Φ_{K3} is indexed by a compact set and so $\lambda_{\Phi_{K3}}$ can chosen as described in [10] to account for a continuum of Lassos.

important (not their coefficients). For the implementation in this paper, a strategy for approximating I_{Φ_K} is adopted where for each $j \leqslant L$, a Lasso regression is run using exactly one test function, $\check{\varphi}_j \in \Phi_K$. The choice of $\check{\varphi}_j$ is made based on being likely to select q_{jL} relative to other $\varphi \in \Phi_K$. Specifically, for each j, $\check{\varphi}_j$ is set to the linear combination of $p_{1K}, ..., p_{KK}$ with highest marginal correlation to q_{jL}. Then the approximation to the first stage model selection step proceeds by using $\check{I}_{\Phi_K} = \bigcup_{j \leqslant L} I_{\check{\varphi}_j(x)}$ in place of I_{Φ_K}. This is also detailed in the appendix.

The formal theory in the subsequent sections proceeds by working with a notion of density of Φ_K within a broader space of approximating functions for $g(x)$. Aside from added generality, working in this manner is helpful since it adds structure to the proofs and it isolates exactly how the density of Φ_K interacts with the final estimation quality for θ_0.

4. Formal Theory

In this section, additional formal conditions are given which guarantee convergence and asymptotic normality of the Post-Nonparametric Double Selection. There are undoubtedly many aspects of the estimation strategy that can be analyzed. These include important choices of tuning parameters and K.

The following definition helps characterize smoothness properties of target function g_0 and approximating functions p^K. Let g be a function on $\mathcal{X}$. Define the Sobolev norm $|g|_d = \sup_{x \in \mathcal{X}} \max_{|c| \leqslant d} |\partial^{|c|} g / \partial x^c|$ where the inner maximum ranges over multi-indeces c.

Assumption 3 (Regularity for p^K)**.** *For each K, there is a nonsingular matrix B_K such that the smallest eigenvalue of the matrix*

$$\Omega_{B_K p^K} = \mathrm{E}\left[B_K p^K(x)(B_K p^K(x))'\right]$$

is bounded uniformly away from zero in K. In addition, there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_{x \in \mathcal{X}} \|B_K p^K(x)\|_2 \leqslant \zeta_0(K)$ and $n^{-1}\zeta_0(K)^2 K \to 0$ as $n \to \infty$.

Assumption 4 (Approximation of g_0)**.** *There is an integer $d \geqslant 0$, a real number $\alpha_{g_0} > 0$, and a sequence of vectors $\beta_{g_0,K}$ which depend on K such that $|g_0 - p^{K\prime}\beta_{g_0,K}|_d = O(K^{-\alpha_{g_0}})$ as $K \to \infty$.*

Assumptions 3 and 4 would be identical to Assumptions 2 and 3 from [45] if there was no conditioning variable z present. These assumptions require that the dictionary p^K has certain regularity and can approximate g_0 at a pre-specified rate. The quantity $\zeta_0(K)$ is dictionary specific, and can be explicitly calculated in certain cases. For instance, [45] gives that $\zeta_0 = O(K^{1/2})$ is possible for B-splines. Note that values of α_{g_0} can be derived for particular d, p^K, and classes of functions containing g_0. [45] also gives explicit calculation of α_{g_0} for the leading cases when $p^K(x)$ are power series and regression splines.

The next assumption quantifies the density of Φ_K within $\mathrm{LinSpan}(p^K)$. In order to do so, define the following. Let

$$\rho(g, \Phi_K) = \inf_{k_g \geqslant 1,\ \eta=(\eta_1,...,\eta_{k_g}) \in \mathbb{R}^{k_g}, \varphi_1, ..., \varphi_{k_g} \in \Phi_K} L^{\alpha_z} \sup_{x \in \mathcal{X}} |g(x) - \varphi(x)| + \|\eta\|_1.$$

Assumption 5 (Density of Φ_K)**.** *Each $\varphi \in \Phi_K$ satisfies* $\text{var}(\varphi(x)) \leqslant 1$. *There is a constant $\alpha_\rho \geqslant 0$ such that*

$$\sup_{\{g \in \text{LinSpan}(p^K):\ \text{var}(g(x)) \leqslant 1\}} \rho(g, \Phi_K) = O(K^{\alpha_\rho}).$$

There is nothing special about the constant 1 in $\text{var}(\varphi(x)) \leqslant 1$. It is mainly a tool for helping describe the density of Φ_K. In addition, as mentioned above, the set selected by Lasso as described in the appendix is invariant to rescaling of the left-hand side variable. As a result, imposing restrictions on $\text{var}(\varphi(x))$ is without loss of generality.

The density assumption is satisfied with $\alpha_\rho = 0$ if the $\Phi_K = \Phi_{K,\text{Span}}$ is used since in that case, $\rho_{\mathcal{Z}}(g, \Phi_K)$ is bounded uniformly in g. On the other hand, the density assumption may only be satisfied with $\alpha_\rho = 1/2$ or higher for the basic $\Phi_{K\,\text{Simple}} = \{p_{1K}(x), ..., p_{KK}(x)\}$ option.

The next assumptions concern sparse approximation properties of $q^L(z)$. Two definitions are necessary before stating the assumption. First, a vector X is called s-sparse if $|\{j : X_j \neq 0\}| \leqslant s$. Next, let π_{q^L} denote the linear projection operator. More precisely, for a square integrable random variable v, $\pi_{q^L} v$ is defined by $\pi_{q^L} v(z) = q^L(z)'\beta_{v,L}$ for $\beta_{v,L}$ such that $\text{E}[(v - q^L(z)'\beta_{v,L})^2]$ is minimized. For functions φ of x such that $\varphi(x)$ is square integrable, write $\beta_{\varphi,L} = \beta_{\varphi(x),L}$.

Assumption 6 (Sparsity)**.** *There is a sequence $s_0 \geqslant 1$ and a constant $\alpha_{\mathcal{Z}} > 0$ such that the following hold.*

1. *There is a sequence of vectors β_{h_0,L,s_0} that are s_0-sparse with support S_0 such that $\sup_{z \in \mathcal{Z}} |h_0(z) - q^L(z)'\beta_{h_0,L,s_0}| = O(L^{-\alpha_{\mathcal{Z}}})$.*
2. *For all $\varphi \in \Phi_K$ there are vectors β_{φ,L,s_0} that are s_0-sparse, all with common support S_0, such that*

$$\sup_{\varphi \in \Phi_K} \sup_{z \in \mathcal{Z}} |\pi_{q^L}\varphi(z) - q^L(z)'\beta_{\varphi,L,s_0}| = O(L^{-\alpha_{\mathcal{Z}}}).$$

Assuming a uniform bound for the sparse approximation error for h_0 is potentially stronger than necessary. At the moment of the writing of the manuscript, the author sees no theoretical obstacle in terms of working under the weaker assumption $n^{-1}\sum_{i=1}^n |h_0(z_i) - q^L(z_i)'\beta_{h_0,L,s_0}|^2 = O_p(L^{-\alpha_{\mathcal{Z}}})$. In addition, the $L^{-\alpha_{\mathcal{Z}}}$ rate is imposed in order to maintain a parallel exposition relative to the $O(K^{-\alpha_{g_0}})$ term. Other rates, for instance $n^{-\alpha_{\mathcal{Z}}}$, can also replace $L^{-\alpha_{\mathcal{Z}}}$, and this is done in [15], [8] and other references.[19] The same comment holds for the sparse approximation conditions for $\varphi \in \Phi_K$.

Several references in the prior econometrics literature work with sparse approximation of the conditional expectation rather than the linear projection. In this context, working with the conditional expectation places a higher burden on the approximating dictionary q^L. In particular, If the conditional expectation of $\varphi(x)$ given z can be approximated using s_0 terms from q^L, then the conditional expectation of $\varphi(x)^2$ may potentially require $O(s_0^2)$ terms to approximate once interactions are taken into account. This potentially requires the dictionary q^L to contain a prohibitively large amount of interaction terms. For this reason, the conditions in this paper are cast in terms of linear projections.[20]

[19] using $n^{-\alpha_{\mathcal{Z}}}$ is only more general if L grows faster than every polynomial of n.

[20] The author sees no theoretical obstructions in terms of applying the same arguments for Lasso bounds in [8] without the conditional expectation assumption. The key ingredient in that

The next assumption imposes limitations on the dependence between x and z. For example, in the case that $\varphi(x) = x$ is an element of $p^K(x)$, this assumption states that the residual variation after a linear regression of x on z is non-vanishing. More generally, the assumption requires that population residual variation after projecting $p_{kK}(x)$ away from z is non-vanishing uniformly K, L. One consequence of Assumption 7 is that constants cannot be freely added to both $g_0(x)$ and $h_0(z)$. This therefore requires the user to enforce a normalization condition like $g_0(0) = 0$ or $\mathrm{E}[g_0(x)] = 0$. The simulation study and empirical illustration below both enforce $g_0(0) = 0$.

Assumption 7 (Identifiability)**.** *For each K and for B_K as in Assumption 3, the matrix $\mathrm{E}[B_K(p^K(x) - \pi_{q^L} p^K(z))(B_K(p^K(x) - \pi_{q^L} p^K(z)))']$ has eigenvalues bounded uniformly away from zero in K, L. In addition, $\sup_{z \in \mathbb{Z}} \|B_K \pi_{q^L} p^K(x)\|_2 \leqslant \zeta_0(K)$.*

The next condition restricts the sample Gram matrix of the second dictionary. A standard condition for nonparametric estimation is that for a dictionary P, the Gram matrix $n^{-1}P'P$ eventually has eigenvalues bounded away from zero uniformly in n with high probability. If $K+L > n$, then the matrix $n^{-1}[PQ]'[PQ]$ will be rank deficient. However, in the high-dimensional setting, to assure good performance of Lasso, it is sufficient to only control certain moduli of continuity of the empirical Gram matrix. There are multiple formalizations of moduli of continuity that are useful in different settings, see [17], [59] for explicit examples. This paper focuses on a simple condition that seems appropriate for econometric applications. In particular, the assumption that only small submatrices of $n^{-1}Q'Q$ have well-behaved eigenvalues will be sufficient for the results that follow. In the sparse setting, it is convenient to define the following sparse eigenvalues of a positive semi-definite matrix M:[21]

$$\kappa_{\min}(m)(M) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|_2^2}, \qquad \kappa_{\max}(m)(M) := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|_2^2}.$$

In this paper, favorable behavior of sparse eigenvalues is taken as a high level condition and the following is imposed.

Assumption 8 (Sparse Eigenvalues)**.** *There is a sequence $s_\kappa = s_\kappa(n)$ such that $s_\kappa \to \infty$ and such that the sparse eigenvalues obey $\kappa_{\min}(s_\kappa)(n^{-1}Q'Q)^{-1} = O(1)$ and $\kappa_{\max}(s_\kappa)(n^{-1}Q'Q) = O(1)$ with probability $1 - o(1)$.*

The assumption requires only that sufficiently small submatrices of the large $p \times p$ empirical Gram matrix $n^{-1}Q'Q$ are well-behaved. This condition seems reasonable and will be sufficient for the results that follow. Informally it states that no small subset of covariates in q^L suffer a multicollinearity problem. They could be shown to hold under more primitive conditions by adapting arguments found in [9] which build upon results in [57] and [49]; see also [48].

---

argument is that expression $\sum_{i=1}^n q^L(z_i)(\pi_{q^L}(z_i) - q^L(z_i)\beta_{\varphi,L})$ stays suitably small. Note this expression is a sum of mean zero independent random variables in the present context.

[21] In the sparse eigenvalue definition, $\| \cdot \|_0$ refers to the number of nonzero components of a vector $(\cdot)$.

Assumption 9 (High-Level Model Selection Performance)**.** *There are constants* α_{I_Φ} *and* α_Φ *and bounds*

1. $|I_{\Phi_K}| \leqslant K^{\alpha_{I_\Phi}} O(s_0)$
2. $\sup_{\varphi \in \Phi_K} \sum_{i=1}^n (q^L(z_i)'(\beta_{\varphi,L,s_0} - \widehat{\beta}_{\varphi,L,\text{Post-Lasso}}))^2 = O(K^{\alpha_\Phi} s_0 \log(L))$
3. $|I_{\text{RF}}| = O(s_0)$
4. $\sum_{i=1}^n (q^L(z_i)'(\beta_{y,L,s_0} - \widehat{\beta}_{y,L,\text{Post-Lasso}}))^2 = O(s_0 \log(L))$

which hold with probability $1 - o(1)$.

The standard Lasso and Post-Lasso estimation rates when there is only one outcome considered are $s_0 \log(L)$ for the sum of the squared prediction errors, and $O(s_0)$ for the number of selected covariates. Therefore, K^{α_Φ} is a uniform measure of the loss of estimation quality stemming from the fact that Lasso estimation is performed on all $\varphi \in \Phi_K$ rather than just on a single outcome. Similarly, $K^{\alpha_{I_\Phi}}$ measures the number of unique j selected in all first stage Lasso estimations. The choice to present high-level assumptions is for generality - so that other model selection techniques can also be applied. However, verification of the high level bounds are available under additional regularity for Lasso estimation.

One reference on performance bounds for a continuum of Lasso estimation steps is [10]. In that paper, the authors provide formal conditions (specifically Assumption 6.1) and prove that Statement 2 of Assumption 9 holds. The bounds in that reference correspond to taking $\alpha_\Phi = 1/2$. An important note is that the conditions in [10] are slightly more stringent since the authors assume that β_{φ,L,s_0} and β_{y,L,s_0} can be taken to approximate the conditional expectation of $\varphi(x)$ and y given z rather than just the linear projection. When $|\Phi_K|$ finite, but grows only polynomially with n and $L > n$, $\alpha_\Phi = 0$ is possible under further regularity conditions.

The main theoretical difficulty in verifying Assumption 9 using primitive conditions is in showing that the size of the set I_{Φ_K} stays suitably small. [10] prove certain performance bounds for a continuum of Lasso estimates under the assumption that $\dim \Phi_K$ is fixed and state that their argument would hold for certain sequences $\dim \Phi_K \to \infty$. [10] also proves that the size of the supports of the Lasso estimates, $|I_{\varphi,L}|$ stay bounded uniformly by a constant multiple of s_0 which does not depend on n or φ. They do not, however, prove that the size of the union $|\cup_{\varphi \in \Phi_K} I_{\varphi,L}|$ remains similarly bounded. Therefore, their results do not imply a the existence of a finite value of α_{I_Φ}. The later bound is required for the analysis of the above proposed estimator. For a finite approximation to $\Phi_{K,\text{Span}}$ (like $\Phi_{K,\text{ Simple}}$), there is no difficulty calculating bounds on the total number of distinct selected terms. This is because under regularity conditions standard in the literature, each $I_{\varphi,L}$ satisfies $|I_{\varphi,L}| \leqslant O(s_0)$ where the implied constants in the $O(s_0)$ terms can be bounded uniformly over $\varphi \in \Phi_K$. In particular, when Φ_K is finite, it is possible to take $K^{\alpha_{I_\Phi}} = |\Phi_K|$. This paper does not derive a bound for $|I_{\Phi_{K,\text{ Span}}}|$ as this would likely lie outside the scope of this project. A valid alternative for which verifiable bounds on the union of selected covariates is possible is to report estimates using

$$\widehat{\Phi}_K = \begin{cases} \Phi_{K,\text{ Simple}} \text{ on the event that } |I_{\Phi_{K,\text{ Span}}}| > t(n) \\ \Phi_{K,\text{ Span}} \text{ otherwise.} \end{cases}$$

for some increasing threshold function t of n.

When $\{g \in \text{LinSpan}(p^K) : \text{var}(g(x)) \leqslant 1\}$ coincides with Φ_K, so that Φ_K is as dense as possible, then Assumption 9 can be weakened in the following.

Assumption 10 (Alternative High-Level Model Selection Performance)**.** *Suppose that $\{g \in LinSpan(p^K) : var(g(x)) \leqslant 1\} \subseteq \Phi_K$. Let ${\Phi_K}' \subset \Phi_K$ be any nonrandom fixed finite subset of at most K elements. There are constants α_{I_Φ} and α_Φ and bounds*

1. $|I_{\Phi_K}| \leqslant K^{\alpha_{I_\Phi}} O(s_0)$
2. $\sup_{\varphi \in {\Phi_K}'} \sum_{i=1}^n (q^L(z_i)'(\beta_{\varphi,L,s_0} - \widehat{\beta}_{\varphi,L,\text{Post-Lasso}}))^2 = O(K^{\alpha_\Phi} s_0 \log(L))$
3. $|I_{\text{RF}}| = O(s_0)$
4. $\sum_{i=1}^n (q^L(z_i)'(\beta_{y,L,s_0} - \widehat{\beta}_{y,L,\text{Post-Lasso}}))^2 = O(s_0 \log(L))$

which hold with probability $1 - o(1)$.

Assumption 10 is weaker than Assumption 9. However, Assumption 9 can be more easily verified with primitive conditions by using finite sets Φ_K.

Statements 2-4 can be attained under standard conditions with $\alpha_\Phi = 0$ provided a penalty adjusting for K different Lasso estimations is used. On the other hand, using a conservative penalty as in [10] for the continuum of Lasso estimations like in $\Phi_{K,\text{Span}}$ would result in $\alpha_\Phi = 1/2$. There is currently no proof that Statement 1 with $\alpha_{I_\Phi} = 0$ and Statement 2 with $\alpha_\Phi = 0$ can hold simultaneously under conditions standard in the econometrics literature.

It is interesting to note that the requirements to satisfy Assumption 10 are essentially pointwise bounds on the predictive performance of a set of Lasso estimations along with a uniform bound on the identity of selected covariates. By contrast, [10] prove uniform bounds on Lasso estimations along with pointwise bounds on the identity of selected covariates. In practice, verification of the Condition 1 in Assumption 10 could be potentially very useful. This would allow the researcher to use a penalty level which is smaller by a factor of $K^{1/2}$, and would ultimately allow more robustness without increasing variability of the final estimator.

For the choice of penalty parameters given in Appendix A for the Span option, Conditions 2-4 of Assumption 10 can be verified under further regularity conditions like those given in [10] or [8] to yield $\alpha_\Phi = 0$. Furthermore, Condition 1 of Assumption 10 can be verified if an option like $\widehat{\Phi}_K$ mentioned on the previous page is used. Most importantly, Assumption 10 serves a plausible high-level model selection condition which is sufficient for proving the results that follow.

The next assumption describes moment conditions needed by applying certain laws of large numbers, for instance for the quantities $n^{-1} \sum_{i=1}^n \varepsilon_i^2 q_{jL}(z_i)^2$.

Assumption 11 (Moment Conditions)**.** *The following moment conditions hold.*

1. $\text{E}[q_{jL}(z)^2 [B_K(p^K(x) - \pi_{q^L} p^K(z))]_k^2]$ *is bounded away from zero uniformly in K, L*
2. $\text{E}[|q_{jL}(z)|^3]$ *is bounded uniformly in L*
3. $\text{E}[q_{jL}(z)^2 \varepsilon^2]$ *is bounded away from zero uniformly in L*
4. $\text{E}[|q_{jL}(z)|^3 |\varepsilon|^3]$ *is bounded uniformly in L.*

The first statement of the assumption may also be seen as a stricter identifiability condition condition on the residual variation $p^K(x) - \pi_{q^L} p^K(z)$. It rules out situations where for instance $x \neq 0 \Leftrightarrow q_{jL}(z) = 0$. Note that $\text{E}[[B_K(p^K(x) - \pi_{q^L} p^K(z))]_k^2] = 1$ is given by the identifiability assumption. No direct assumption

is needed about the corresponding third moment $\mathrm{E}[[B_K(p^K(x) - \pi_{q^L} p^K(z))]_k^3] = 1$ since instead a reference to the bound $\zeta_0(K)$ is used.

The final assumption before the statement of Theorem 1 are rate conditions.

Assumption 12 (Rate Conditions)**.** *The following rate conditions hold.*

1. $s_0 K^{\alpha_{I_\Phi}} = o(s_\kappa)$
2. $\log(KL) = o(\zeta_0(K)^{-1} n^{1/3})$
3. $L^{-\alpha_z} n^{1/2} K^{-1/2} \zeta_0(K) = O(1)$
4. $L^{-2\alpha_z} K^{2\alpha_\rho}(K^{1/2} n^{1/2} \zeta_0(K)^{-1} + n^{1/2} + K \log(L)^{1/2} \zeta_0(K)^{-2}) = O(1)$
5. $n^{-1/2} K^{1/2} s_0 \log(L) \zeta_0(K)^{-1}(K^{2\alpha_\rho + \alpha_\Phi} + K^{\alpha_\rho + \alpha_\Phi/2 + \alpha_{I_\Phi}/2}) = O(1)$
6. $n^{-1/2} s_0^{1/2} \log(L)(K^{2\alpha_\rho + \alpha_\Phi} s_0^{1/2} + K^{\alpha_{I_\Phi}/2}) = O(1)$.

The first statement ensures that the sparse eigenvalues remain well-behaved in the with high probability over sets whose size is larger that the selected covariates. The second statement is used in conjunction with the above moment conditions to allow the use of moderate deviation bounds following [37]. The third and fourth conditions are assumption on the sparse approximation error for $q^L(z)$. The final two assumptions restrict the size of s_0 and K and quantities depending on α_ρ, α_Φ, and α_{I_Φ} relative to n. These assumptions can be unraveled for certain choices of dictionaries. For example, as was noted above and by [45], for B-splines, $\zeta_0(K)$ can be taken to be $O(K^{1/2})$. Using the simple option gives $\alpha_\rho = 1$, $\alpha_\Phi = 0$ and $\alpha_{I_\Phi} = 1$. Then the conditions can be reduced to $L^{-\alpha_z} n^{1/2} = O(1)$, $L^{-2\alpha_z} K^2 n^{1/2} = O(1)$, $n^{-1/2} K^2 s_0 \log(L) = O(1)$.

The first result is a preliminary result which gives bounds on convergence rates for the estimator $\widehat{g}$. They are used in the course of the proof of Theorem 1 below, the main inferential result of this paper. The proposition is a direct analogue of the rates given in Theorem 1 of [45] which considers estimation of a conditional expectation g_0 without model selection over a conditioning set. The rates obtained in Proposition 1 match the rates in [45]. To state it, let F_0 be the distribution function of the random variable x. In addition, let $\zeta_d(K) = \max_{|c| \leqslant d} \sup_{x \in \mathcal{X}} \|\partial^{|c|} B_K p^K(x)/\partial x^c\|_2$.

Theorem 1. *Under Assumptions 1-8, 9 or 10, and 11-12, the Post-Nonparametric Double Selection estimate $\widehat{g}$ for the function g_0 satisfies the following bounds.*

$$\int (\widehat{g}(x) - g_0(x))^2 dF_0(x) = O_p(n^{-1}K + K^{-2\alpha_{g_0}}).$$

$$|\widehat{g} - g_0|_d = O_p(n^{-1/2}\zeta_d(K) K^{1/2} + K^{-\alpha_{g_0}}).$$

The next formal results concern inference for $\theta_0 = a(g_0)$. Recall that θ_0, is estimated by $\widehat{\theta} = a(\widehat{g})$ and inference is conducted via the estimator $\widehat{V}$ as described in earlier sections.

Assumption 13 (Moments for Asymptotic Normality)**.** $\mathrm{E}[\varepsilon^{4+\delta}|x, z]$ *is bounded for some* $\delta > 0$. $\mathrm{var}(\varepsilon|x, z)$ *is bounded away from zero.*

Note that the conditions in [45] require only that $\mathrm{E}[\varepsilon^4|x, z]$ is bounded. The strengthened condition is needed for consistent variance estimation, in order to construct a bound on the quantity $\max_{i \leqslant n} \varepsilon_i^2$.

The following assumptions on the functional a are imposed. They are regularity assumptions that imply that a attains a certain degree of smoothness. For example, they imply that a is Fréchet differentiable.

Assumption 14 (Differentiability for a)**.** *The real valued functional $a(g) \in \mathbb{R}$ is either linear or the following conditions hold. $n^{-1}\zeta_d(K)^4K^2 \to 0$. There is a linear function $D(g;\breve{g})$ that is linear in g and such that for some constants $C, \nu > 0$ and all $\breve{g}, \breve{\breve{g}}$ with $|g - g_0|_d < \nu$, $|\breve{g} - g_0|_d < \nu$, $|\breve{\breve{g}} - g_0|_d < \nu$, it holds that $|a(g) - a(\breve{g}) - D(g - \breve{g}; \breve{g})| \leqslant C(|g - \breve{g}|_d)^2$ and $|D(g;\breve{g}) - D(g;\breve{\breve{g}})| \leqslant C|g|_d|\breve{g} - \breve{\breve{g}}|_d$.*

The function D is related to the functional derivative of a. The following assumption imposes further regularity on the continuity of the derivative. For shorthand, let $D(g) = D(g; g_0)$.

The next rate conditions is used to ensure that estimates are undersmoothed. The rate condition ensures that the estimation bias, which is heuristically captured by $K^{-\alpha_{g_0}}$, converges to zero faster than the estimation standard error.

Assumption 15 (Undersmoothing Rate Condition)**.** $n^{1/2}K^{-\alpha_{g_0}} = o(1)$.

The next rate condition is used in order to bound quantities appearing in the proof of Theorem 2. As was demonstrated in the case of Assumption 12, the rate conditions can be unraveled for certain choices of K, p^K, and Φ_K.

Assumption 16 (Rate Conditions for Asymptotic Normality)**.**

1. $L^{-2\alpha_z}K^{2\alpha_\rho}(\zeta_0(K)K + \zeta_0(K)^4K^{1-2\alpha_\rho} + K\log(L) + n^{1/2}) = o(1)$
2. $n^{-1}s_0\zeta_0(K)\log(L)(K^{1+2\alpha_\rho+\alpha_\Phi} + K^{1+\alpha_\rho+\alpha_\Phi/2+\alpha_{I_\Phi}/2}) = o(1)$
3. $n^{-1}s_0^2\log(L)^2(K^{4\alpha_\rho+2\alpha_\Phi} + K^{2\alpha_\rho+\alpha_\Phi+\alpha_{I_\Phi}}) = o(1)$
4. $s_0K^{\alpha_{I_\Phi}}\left(n^{-1/2}\zeta_0(K)K^{1/2} + K^{-\alpha_{g_0}}\right) = o(1)$
5. $n^{2/(4+\delta)}\zeta_0(K)n^{-1/2}K^{1/2} = o(1)$.

The final two conditions divide the cases considered into two classes. The first class (covered by Assumption 17) are functionals which fail to be mean-square differentiable and therefore cannot be estimated at the parametric $n^{1/2}$ rate. The second class (covered by Assumption 18) does attain the $n^{1/2}$ rate. One example with the functional of interest is evaluation of g at a point x_0: $a(g) = g(x_0)$. In this case, a fails to be estimated that the parametric $n^{1/2}$ rate in general circumstances. A second example is the weighted average derivative $a(g) = \int w(x)\partial g(x)/\partial x$ for a weight function w which satisfies regularity conditions. The Assumption 18 holds if w is differentiable, vanishes outside a compact set, and the density of x is bounded away from zero wherever w is positive. In this case, $a(g) = \mathrm{E}[\psi(x)g(x)]$ for $\psi(x) = -\phi(x)^{-1}\partial w(x)/\partial x$ by a change of variables provided that x is continuously distributed with non vanishing density ϕ. These are one possible set of sufficient conditions under which the weighted average derivative does achieve $\sqrt{n}$-consistency.

Assumption 17 (Regularity for a in Absence of Mean-Square Differentiability)**.** *There is a constant $\bar{C} > 0$ such that $|D(g)| \leqslant \bar{C}|g|_d$. There is $\bar{\beta}$ dependent on K such that for $\bar{g}(x) = p(x)^{K\prime}\bar{\beta}$, it holds that $\mathrm{E}[\bar{g}(x)^2] \to 0$ and $D(\bar{g}) \geqslant \bar{C} > 0$.*

Assumption 18 (Conditions for $n^{1/2}$-Consistency)**.** *There is $\psi(x)$ such that $\mathrm{E}[\psi(x)^2]$ finite and nonzero and such that $D(g) = \mathrm{E}[\psi(x)g(x)]$ and $D(p_{kK}) = \mathrm{E}[\psi(x)p_{kK}(x)]$ for every k. There is $\breve{\beta}$ such that $\mathrm{E}[(\psi(x) - p(x)^{K\prime}\breve{\beta})^2] \to 0$. Finally, the matrix $\bar{V} = \mathrm{E}[\psi(x)^2 var(y|x,z)]$ is finite and nonzero.*

Theorem 2 now establishes the validity of standard inference procedure after model selection as well as validity of the plug in variance estimator.

Theorem 2. *Under Assumptions 1-8, 9 or 10, 11-17, the Post-Nonparametric Double Selection estimate for the function θ_0 satisfies*

$$\widehat{\theta} = \theta_0 + O_p(n^{-1/2}\zeta_d(K)).$$

In addition,

$$n^{1/2}V^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0,1) \quad \textit{and}$$
$$n^{1/2}\widehat{V}^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0,1).$$

Under Assumptions 1-8, 9 or 10, 11-16, and 18,

$$n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, \bar{V}) \quad \textit{and} \quad V - \bar{V} \xrightarrow{p} 0.$$

5. Simulation study

The results stated in the previous section suggest that Post-Nonparametric Double Selection series estimation should exhibit good inferential properties for additively separable conditional expectation models when the sample size n is large. The following simulation study is conducted in order to illustrate the implementation and study the performance of the outlined procedure.

The simulation study is divided into two parts. The first part compares several alternative estimators to Post-Nonparametric Double Selection. The second part compares several Post-Nonparametric Double Selection estimates using different choices for Φ_K. This part demonstrates finite sample benefits from using the Span option relative to the direct generalization of Post-Double Selection estimation (ie. using the Simple option).

The following process generates the data in each simulation.

$$y = g_0(x) + h_0(z) + \varepsilon$$

$$\begin{aligned}
&g_0(x) = 10\sin(0.1x) - 0.5\sin(4\pi x 4^{-x^2}) \\
&h_0(z) = z'\beta_{h_0,L},\ \beta_{h_0,L,j} = -0.5\cdot(-0.65)^{j-1}\mathbf{1}_{j\leqslant s_0} \\
&z_j \sim \mathrm{N}(0,1), \quad \mathrm{corr}(z_{j_1}, z_{j_2}) = 0.25^{|j_1 - j_2|} \\
&\varepsilon \sim \mathrm{N}(0,1) \\
&x = 0.15v + 0.0375 - 3.75(\mathrm{stair}(z'\gamma_0 + v) + 0.375)F_{\mathrm{N}(0,1)}(10z_{s_0})... \\
&\qquad + 3.75(\mathrm{stair}(0.5(z'\gamma_0 + v)|z'\gamma_0 + v|^{0.25})(1 - F_{\mathrm{N}(0,1)}(10z_{s_0})) \\
&\gamma_{0,L,j} = -1.5(-0.75)^{j-1}\mathbf{1}_{j\leqslant s_0} \\
&v \sim \mathrm{N}(0,1) \\
&\mathrm{stair}(\ \cdot\ ) = 0.25\frac{\tanh(12(\ \cdot\ )/2.5) - 12\lfloor(\ \cdot\ )/2.5\rfloor - 6}{2\tanh(6) + 0.5 + \lfloor(\ \cdot\ )/2.5\rfloor}.
\end{aligned}$$

The study performs simulations for $n \in \{100, 150, ..., 500\}$. Two settings for the parameter L are considered: $L = n/2$ and $L = 2n$. Finally, the sparsity level is set to $s_0 = 6$. Within each data generating process, 1000 simulation replications are performed.

The data generating process is quite complicated. It is designed in order to create correlations between the covariates z and various transformations of x. This allows the data generating process to highlight many different statistical problems which can arise using Nonparametric-Post Double Selection and alternative estimation techniques all in one simulation study. Despite the complicated formulas for the joint distribution of x and z, their realizations appear natural. Scatter plots of one sample of $n = 500$ showing the respective bivariate distributions between $z_1, ..., z_6$ and x are provided in Figure 6. Figure 5 provides a picture of the graph of g_0.

The simulations evaluate estimation of g_0 and of θ_0 defined by

$$\theta_0 = \mathrm{E}[g_0'(x)].$$

In order to avoid further complications, for each replication, the expectation and thus true θ_0 are calculated against the empirical distribution of x within that simulation replication.[22]

The first part of the simulation study considers the performances of five estimators[23] for g_0 and θ_0. Each estimator is a reduced series estimator based on initial dictionaries consisting of a cubic spline expansion $p^K(x)$ for $g_0(x)$ and a linear expansion $q^L(z) = z$ for $h_0(z)$.

1. **Oracle.** Estimator 1 is infeasible and sets $\tilde{q}(z) = (z_1, .., z_{s_0})$. This estimator serves as a benchmark for comparison to estimates in which the correct support is known.
2. **Span Post-Nonparametric Double.** Estimator 2 selects $\tilde{q}(z)$ using Post-Nonparametric Double Selection with Φ_k given by the Span option, as described in this paper.
3. **Naive.** Estimator 3 selects $\tilde{q}(z)$ in one model selection step by performing Lasso of y on $q^L(z)$.
4. **OLS.** Estimator 4 uses $\tilde{q}(z) = z$. In other words, this estimator does not reduce the dictionary. This estimation strategy is only calculated provided $L < n$.
5. **Targeted Undersmoothing.** Estimator 5 implements an alternative inferential procedure for dense functionals of high-dimensional parameters; TU(1). This procedure was proposed in [30] and is described further below.

[22]Another possibility is to calculate against the population expectation of x. Under the assumption that the researcher knows the population distribution of x, this causes no further complication. If the distribution of x is unknown and estimated, this must however be taken into account.

[23]There are likely other sensible estimators beyond the 5 considered in the simulation section. As pointed out by an anonymous reviewer, such estimators may include propensity score matching on a continuous variable. Though such an approach may work well, the context here is not exactly the same as usually seen in propensity score matching. In particular, the assumptions here do not require unconfoundedness conditions. In addition, propensity score techniques are most commonly applied to discrete treatement variables. There is some work on propensity score matching with a continuous treatment; for example, see [35], who require the estimation of the conditional density of treatment. In the high-dimensional setting, estimating the conditional density of x given z would likely introduce complications beyond the scope of this paper.

Detailed implementation descriptions are provided in Appendix A. For each of the above estimators, the choice of $p^K(x)$ is made using a data-dependent rule. First, an initial dictionary reduction $q^{\text{initial}}(z)$ is selected. For Oracle, $q^{\text{initial}}(z) = (z_1, .., z_{s_0})$. For the Span Post-Nonparametric Double and Naive estimators, $q^{\text{initial}}(z)$ is based on Lasso of y on $q^L(z)$ as implemented in Appendix A. For OLS, $q^{\text{initial}}(z) = z$. Next, BIC is used to choose a B-spline expansion $p^K(x)$.

Comparison of estimators 1-4 is standard in the post-model selection econometrics literature. The oracle estimator should be seen as a benchmark which is known to provide good estimates if the true set, S_0, was known. The Naive estimator is expected to perform poorly since it is not a uniformly valid estimator and susceptible to size-distortions arising from model selection mistakes. OLS is expected to perform poorly due to potential problems related to overfitting.

Estimator 5 is a procedure called *Targeted Undersmoothing* which looks to correct distortions in inference from model selection mistakes. Targeted Undersmoothing appends covariates which significantly affect the value of the functional $\widehat{\theta} = a(\widehat{g})$ to an initially selected model (see [30]). It is appropriate for functionals of high-dimensional models which depend on a growing number of parameters (dense functionals) and is therefore a potentially sensible procedure for inference for θ_0. This estimator is detailed further in Appendix A.

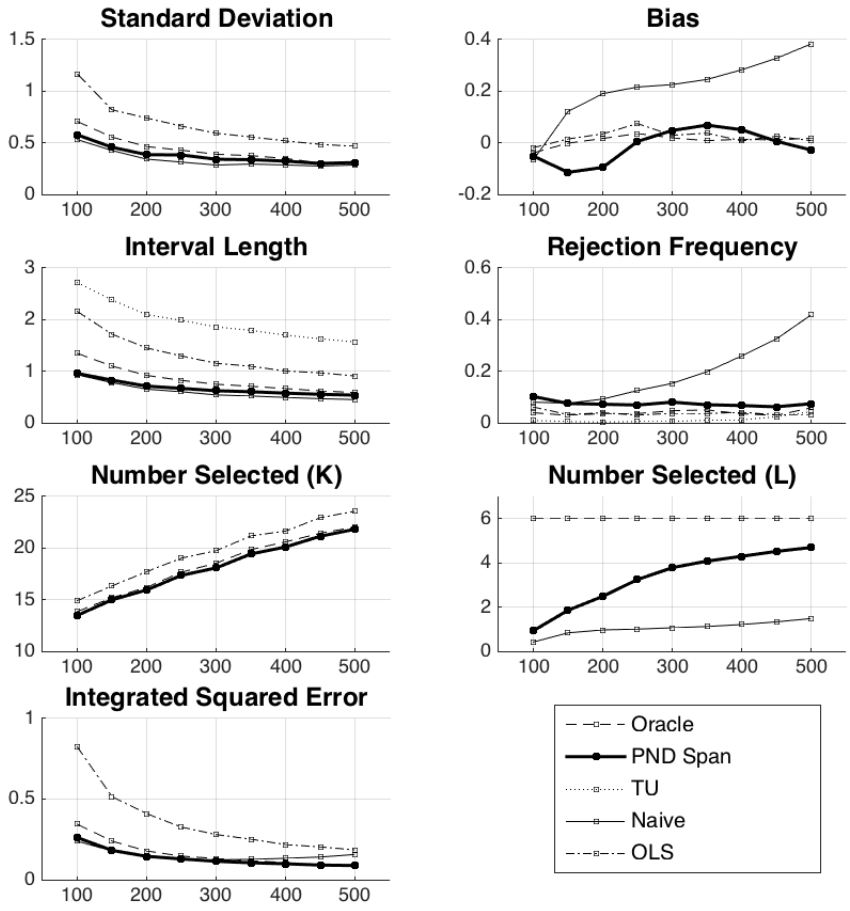
The simulation results report several quantities which measure the performance of each estimator. The results report standard deviation of the estimates $\widehat{\theta}$, bias of the estimates for θ_0, confidence interval length for estimates for θ_0, rejection frequencies under the null for θ_0 at the 5% level, mean number of series terms K used, mean number of series terms selected from the original L, and integrated squared error for g_0. The simulation results are reported in Figure 1 for $L = n/2$ and Figure 2 for $L = 2n$. The figures display the above mentioned simulation results for each $n = 100, ..., 500$ with n changing over the horizontal axis.[24] Note also that across some of the estimators, some of the reported quantities will be identical. For example, the point estimates for TU are identical to the Naive point estimates. The selected K is identical for the Naive estimates as well as the Post-Nonparametric Double Selection estimates.

In all of the simulations, the Post-Nonparametric Double Selection estimates behave similarly to the Oracle estimates. The OLS estimates have wide confidence intervals relative to the Post-Nonparametric Double Selection estimation, but have similar coverage properties. The final estimator, Targeted Undersmoothing (TU), is conservative in terms of coverage, with substantially larger intervals in every case.

On the other hand, the Naive estimator has poor coverage properties. For the Naive estimator, after failing to control for the correct covariates, the increase in K leads to an increasing bias. This highlights the fact that simply producing undersmoothed estimates of g_0 by increasing K may not be adequate for reducing bias and making quality statistical inference possible in the high-dimensional setting.

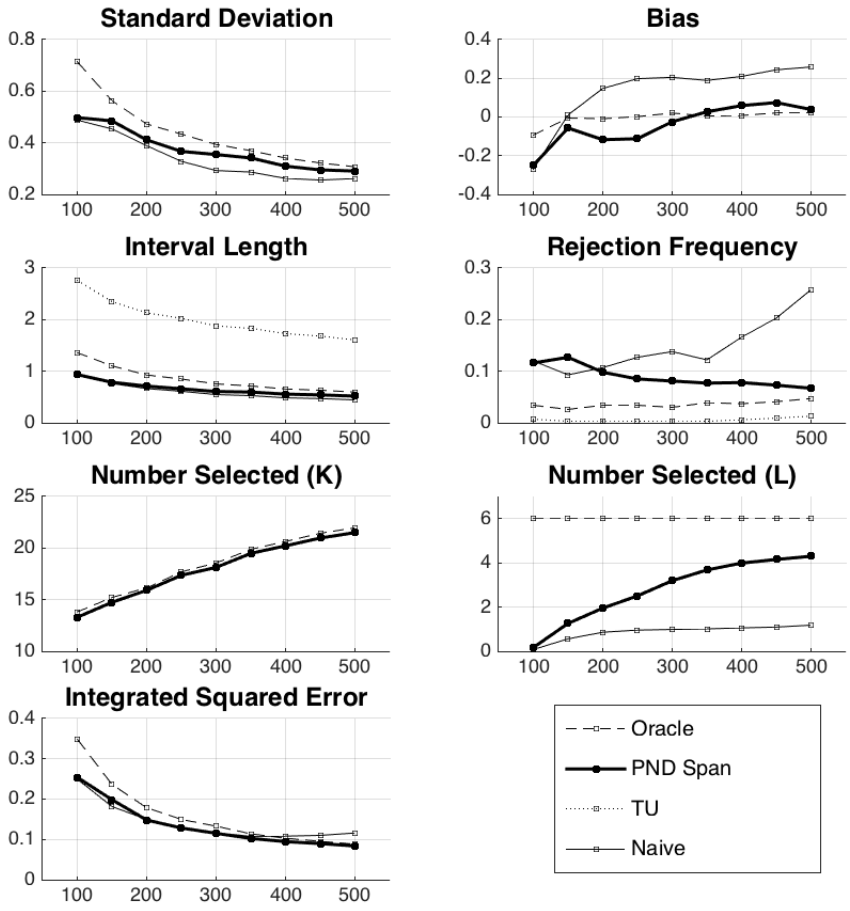
[24]Note that since s_0, the magnitude of coefficients $\beta_{h_0,L}$ and the joint distribution between relevant covariates are all fixed in the simulations as $n \to \infty$. Therefore, for sufficiently large n, all relevant covariates would be identified with high probability, and all of the post-model selection estimators would perform similarly. This simulation study therefore is identifying differences in finite sample performance.

FIGURE 1. Simulation Results



This figure presents simulation results for the estimation of g_0 and θ_0 in the cases $n = 100, 150, ..., 500$ with $s_0 = 6$ and $L = n/2$ according to the data generating process described in the text. Estimates are presented for the five estimators, Oracle, Post-Nonparametric Double (PND Span), Naive, OLS and Targeted Undersmoothing (TU) as described in the text. The first plot shows standard deviation of the respective estimates for θ_0. The second plot shows bias of the estimates for θ_0. The third plot shows confidence interval length for estimates for θ_0. The fourth plot shows rejection frequencies under the null for θ_0 for a 5% level test. The fifth plot shows the mean number of series terms K used. The sixth plot shows the mean number of series terms from L selected. The seventh plot shows root mean integrated squared error for g_0. Figures are based on 1000 simulation replications. n is always indexed by the horizontal axis.

FIGURE 2. Simulation Results



This figure presents simulation results for the estimation of g_0 and θ_0 in the cases $n = 100, 150, ..., 500$ with $s_0 = 6$ and $L = 2n$ according to the data generating process described in the text. Estimates are presented for the four estimators, Oracle, Post-Nonparametric Double (PND Span), Naive, and Targeted Undersmoothing (TU) as described in the text. The first plot shows standard deviation of the respective estimates for θ_0. The second plot shows bias of the estimates for θ_0. The third plot shows confidence interval length for estimates for θ_0. The fourth plot shows rejection frequencies under the null for θ_0 for a 5% level test. The fifth plot shows the mean number of series terms K used. The sixth plot shows the mean number of series terms from L selected. The seventh plot shows root mean integrated squared error for g_0. In each plot, the horizontal axis denotes sample size n. Figures are based on 1000 simulation replications. n is always indexed by the horizontal axis.

The second part of the simulation study compares four Post-Nonparametric Double Selection estimators which use different specifications for Φ_K.

1. **Span Post-Nonparametric Double.** Estimator 1 is identical to the Span Post-Nonparametric Double estimator in the first part of the simulation.
2. **Conservative Span Post-Nonparametric Double.** Estimator 2 uses p^K and Φ_K as in the Span option, but in the decomposition $\Phi_{K,\text{Span}} = \Phi_{K1} \cup \Phi_{K2} \cup \Phi_{K3}$, the penalty applied to Φ_{K3} is more conservative, explicitly aimed at achieve Lasso performance bounds which hold uniformly over all of Φ_K.
3. **Simple Post-Nonparametric Double.** Estimator 3 uses p^K as in the Span, but uses $\Phi_K = \Phi_{K,\text{ Simple}}$.
4. **Alternative Spline Basis Simple Post-Nonparametric Double.** Estimator 4 uses a different basis for selection. A QR decomposition is applied to P in order to obtain orthonormal columns. Next, $\Phi_K = \Phi_{K,\text{ Simple}}$ is used on the new orthogonalized data. Importantly, the new P spans the same K-dimensional linear space in $\mathbb{R}^n$ as in the 3 previous estimators.

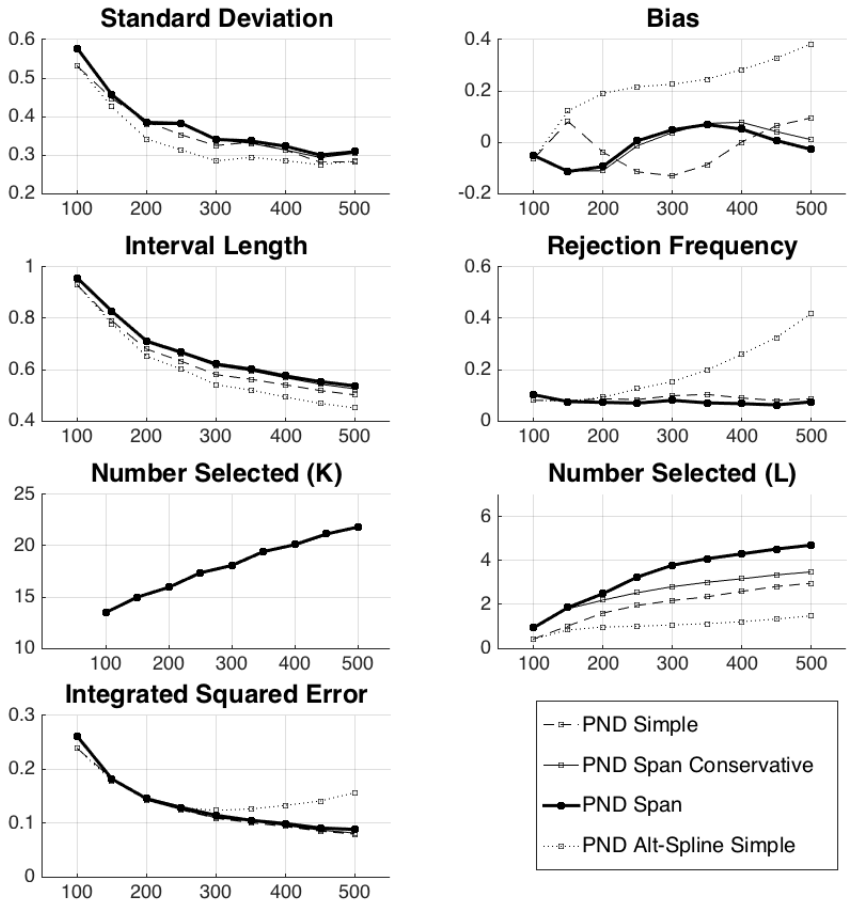
The estimates for the second part of the simulation are presented in Figures 3-4. Note that all estimators are identical with regards to K, hence only one curve is visible in the corresponding plots. In addition, the Conservative Span and Span estimators have very similar performance in terms of standard deviation, bias, interval length, rejection frequency, and integrated squared error. The two estimators are practically indistinguishable except in terms of the number of elements of q^L they select. They do not give numerically identical estimates or confidence intervals. However, their differences are too small to be seen in Figures 3-4.

There are noticeable differences in the performance of the estimators. The Span option is able to identify the highest number of relevant covariates, followed by the Conservative Span option, the Simple option, and the Alternative Spline Basis Simple option. The Span, Conservative Span, and Simple Post-Nonparametric Double Selection procedures exhibit favorable finite sample properties for this data generating process. In particular, for those estimators, the calculated rejection frequencies move towards 5% as n increases.

By contrast, the Alternative Spline Basis Simple Post-Nonparametric Double Selection procedure has very poor finite sample performance. It is unlikely that the projection of the new orthogonalized basis onto q^L has a good sparse representation. This causes increased model selection mistakes in the first stage. Unlike in the partially linear model, these mistakes can accumulate to cause more severe bias since the number of first stage selection steps is growing with K. Note that the Alternative Spline Basis estimator has similar performance to the Naive estimator in the first part of the simulation study.

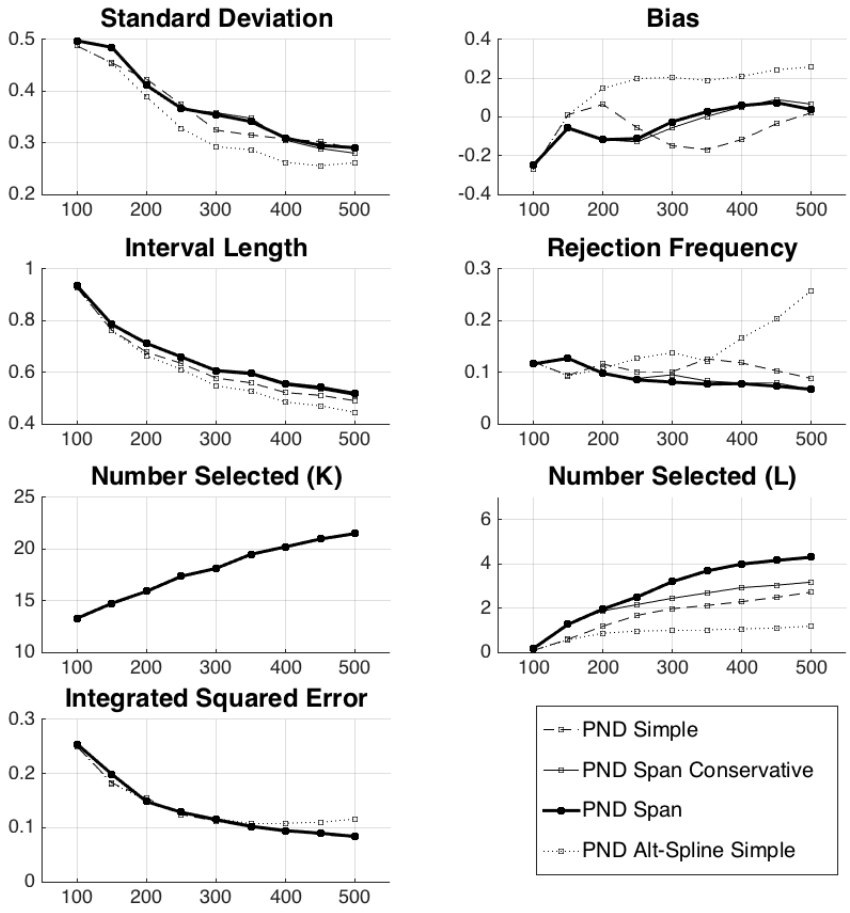
The Span and the Conservative Span options offer an opportunity to potentially add additional robustness. These options select more variables than the Simple option. There is no evidence from this simulation study that using the Span option over-selects conditioning variables to the extent that rejection frequencies become severely distorted or variability increases to an undesirable level.

FIGURE 3. Simulation Results



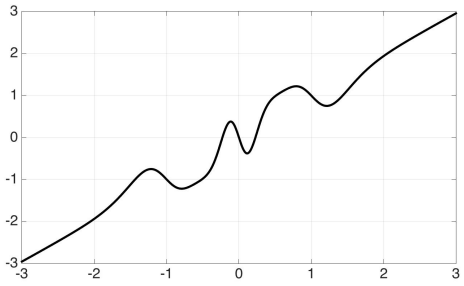
This figure presents simulation results for the estimation of g_0 and θ_0 in the cases $n = 100, 150, ..., 500$ with $s_0 = 6$ and $L = n/2$ according to the data generating process described in the text. Estimates are presented for four Post-Nonparametric Double Selection (PND) estimators, Simple, Span, Conservative Span, and Alternative Spline Simple as described in the text. The first plot shows standard deviation of the respective estimates for θ_0. The second plot shows bias of the estimates for θ_0. The third plot shows confidence interval length for estimates for θ_0. The fourth plot shows rejection frequencies under the null for θ_0 for a 5% level test. The fifth plot shows the mean number of series terms K used. The sixth plot shows the mean number of series terms from L selected. The seventh plot shows root mean integrated squared error for g_0. In each plot, the horizontal axis denotes sample size n. Figures are based on 1000 simulation replications. n is always indexed by the horizontal axis.

FIGURE 4. Simulation Results



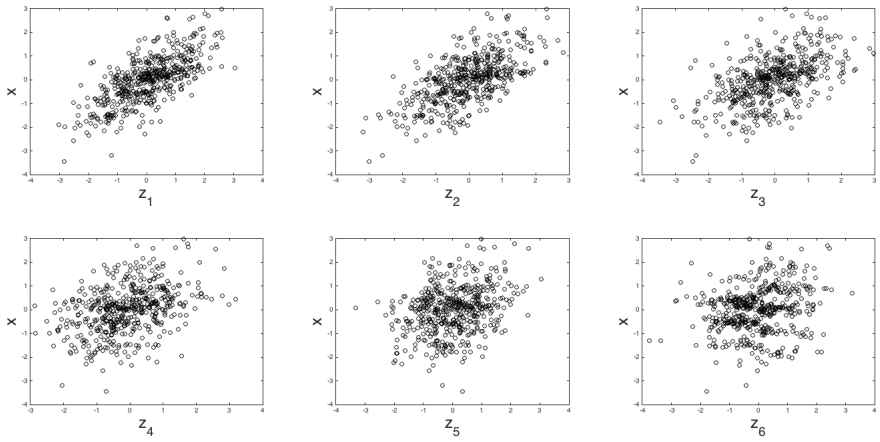
This figure presents simulation results for the estimation of g_0 and θ_0 in the cases $n = 100, 150, ..., 500$ with $s_0 = 6$ and $L = 2n$ according to the data generating process described in the text. Estimates are presented for the four Post-Nonparametric Double Selection (PND) estimators, Simple, Span, Conservative Span, and Alternative Spline Simple as described in the text. The first plot shows standard deviation of the respective estimates for θ_0. The second plot shows bias of the estimates for θ_0. The third plot shows confidence interval length for estimates for θ_0. The fourth plot shows rejection frequencies under the null for θ_0 for a 5% level test. The fifth plot shows the mean number of series terms K used. The sixth plot shows the mean number of series terms from L selected. The seventh plot shows root mean integrated squared error for g_0. In each plot, the horizontal axis denotes sample size n. Figures are based on 1000 simulation replications. n is always indexed by the horizontal axis.

FIGURE 5. Simulation study: g_0



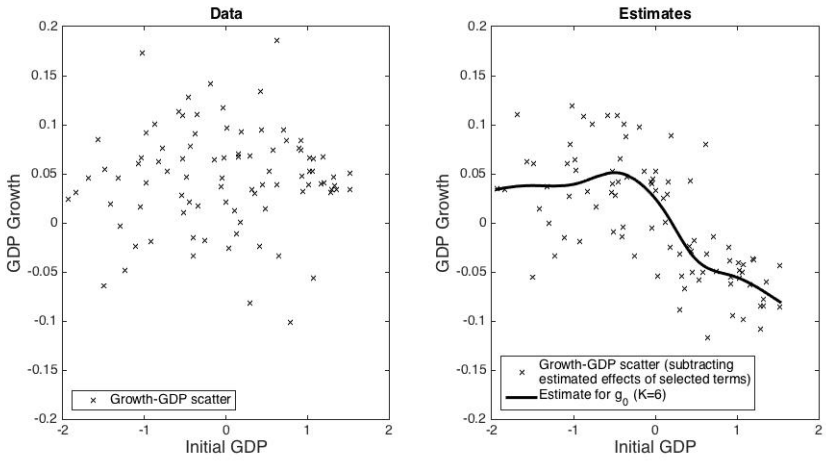
This figure depicts the function g_0 used in the simulation study.

FIGURE 6. Simulation study: joint covariate distribution



This figure depicts the joint distribution between x and the first $s_0 = 6$ covariates as described in the above text. The plots are generated by one sample of size $n = 500$.

FIGURE 7. GDP Growth Results



6. Empirical Example: GDP Growth

This section applies Post-Nonparametric Double Selection to an international economic growth example. The data comes from the Barro and Lee [6] dataset which contains a panel of 138 countries for the period of 1960 to 1985. This example was also considered in [13], who apply Lasso techniques in the context of a high-dimensional linear model for the purpose of locating important variables which are predictive of GDP growth rates. This considers growth in GDP per capita as a dependent variable y for the period 1965-85. The growth rate in GDP over a period from t_1 to t_2 is commonly defined as $\log(\text{GDP}_{t_2}/\text{GDP}_{t_1})$.

Studying the factors that influence growth in GDP is a problem of central importance in economics. A difficulty with studying this problem empirically on a cross-country level is that the number of observations is limited by the total number of countries. At the same time, the number of potential factors which influence GDP growth can be large. This leads naturally to the need to regularize econometric estimation on any data on a cross-section of countries. This example specifically studies the relation between initial GDP level and subsequent GDP growth in the presence of a large number of other determinants of GDP growth. The interest in studying this particular question is in testing the fundamental macroeconomic theory of convergence. Convergence predicts that countries with high initial GDP will show lower levels of GDP growth, and conversely countries with low initial GDP will show higher levels of GDP growth. There are many references for assumptions which imply such convergence. See [1] and references therein.

This analysis considers a model with $p = 62$ covariates, which allows for a total of $n = 90$ complete observations. Since p is comparably large relative to n, dimension reduction in this setting is necessary. The goal here is to select a subset of these covariates and briefly compare the resulting to predictions made in the growth literature (see [7], [6]). [24] and [6] contain complete definitions and discussion of each of these variables. The estimated model is given by the specification

$$y_i = g_0(\ \log(\text{GDP}_i) - \overline{\log(\text{GDP}_i)}\ ) + h_0(z_i) + \varepsilon_i$$

where $\overline{\log(\text{GDP}_i)}$ denotes the sample mean. The observed covariates enter linearly, so that the expansion $h_0(z_i) = z_{i1}\beta_1 + ... + z_{i62}\beta_{62}$ is assumed.

The estimation is performed using cubic splines as detailed in Appendix A. g_0 is normalized so that $g_0(0) = 0$. Estimates of several density-weighted average derivatives of the effect of initial GDP on GDP growth are constructed using Post-Nonparametric Double Selection and are presented in Table 1. In addition, a scatter plot of the primary variables of interest as well as an estimate of g_0 are shown in Figure 7.

A nonlinear specification for g_0 allows testing of several hypotheses related to the convergence of GDP. These include the hypothesis that conditional convergence can depend on initial GDP. This is related to the idea of a poverty trap where countries with smaller initial GDP exhibit less convergence (ie. the relationship between initial GDP and GDP growth may be locally flat; see the reference text [1] for additional background and details.) Conditional convergence could also imply that at the high end of the initial GDP distribution, GDP growth is locally flat. The existence of conditional convergence based on initial GDP can be tested by using a nonlinear specification for g_0. In order to study the overall convergence, the data is divided into quartiles. An average derivative is then estimated within each quartile. In addition, an overall average derivative is estimated over the support of all initial GDP observations. The respective average derivatives are then compared.

Estimates based on Post-Nonparametric Double Selection are presented in Table 1. The estimate for the overall weighted average derivative is -0.042 (std. err. = 0.014, p = 0.003).[25] The estimate is negative and statistically significant. This result is consistent with convergence theory. In addition, the density-weighted average derivative is calculated for various smaller ranges of initial GDP. The empirical distribution of initial GDP is divided into quartiles. Estimates for the weighted average derivatives are calculated within each quartile. The estimated average derivatives are 0.013 (std. err. = 0.022, p = 0.568) for Q_1, -0.042 (std. err. = 0.025, p = 0.087) for Q_2, -0.100 (std. err. = 0.041, p = 0.015) for Q_3, -0.062 (std. err. = 0.027, p = 0.022) for Q_4. The test of the hypothesis that the average derivative in Q_1 is equal to the average derivative over Q_2, Q_3, Q_4 rejects the null at the 5% level (p = 0.009, t stat. = -2.632). The test of the hypothesis that the average derivative in Q_4 is equal to the average derivative over Q_1, Q_2, Q_3 fails to reject the null at the 5% level (p = 0.908, t stat. = -0.115). The overall average derivative estimate is negative and statistically significant. These estimates also agree with and thus support the previous findings reported in [24], [6], [7], which relied on ad-hoc reasoning for covariate selection. In addition, the analysis supports the claim that conditional convergence is nonlinear in initial GDP, being flatter for countries with lower initial GDP.

[25] p-values are calculated against a two-sided alternative for the null that the average derivative is 0.

TABLE 1. Estimation Results for GDP Example.

	Estimates	(S. E.)	[95% C. I.]	
Average Derivative				
	-0.042	(0.014)	[-0.070	-0.014]
Quartile-Specific Average Derivative				
Q_1	-0.013	(0.022)	[-0.031	0.056]
Q_2	-0.042	(0.025)	[-0.090	0.006]
Q_3	-0.100	(0.041)	[-0.182	-0.019]
Q_4	-0.062	(0.027)	[-0.115	-0.009]

Additional Selected Variables

- Life expectancy
- Average schooling years in female population over age 25
- Infant mortality rate
- Female gross enrollment ratio for secondary education
- Male gross enrollment ratio for secondary education
- Total fertility rate
- Population proportion under 15

Additional Hypothesis Tests

H_0 : ave. deriv. (Q_1) = ave. deriv. (Q_2-Q_4)
p-value = .009
t-statistic = -2.632
H_0 : ave. deriv. (Q_4) = ave. deriv. (Q_1-Q_3)
p-value = .9084
t-statistic = 0.115

Note. Post-Nonparametric Double Selection estimates with B-spline basis. $\widehat{K} = 7$.

7. CONCLUSION

This paper considers the problem of selecting a conditioning set in the context of nonparametric regression. Convergence rates and inference results are provided for series estimators of a primary component of interest in additively separable models with high-dimensional conditioning information. The finite sample performance of several Post-Nonparametric Double Selection estimators are evaluated in a simulation study. Overall, the proposed Span option has good estimation and inferential properties in the data generating processes considered.

Appendix A. Implementation Details

A.1. **Lasso Implementation Details.**

A.1.1. *Lasso implementation given penalty λ.* In every case, penalty loadings ℓ_j are chosen as described in [8] with one small modification. The procedure suggested in [8] requires an initial penalty loadings which are constructed using initial estimates of regression residuals. Their suggestion is to use $\widehat{\varepsilon}_i^{\text{initial}} = y_i$ followed by an iterative procedure. Here, instead, $\widehat{\varepsilon}_i^{\text{initial}}$ are taken as the linear regression residuals after regressing the outcome v on the 5 most marginally correlated q_{jL}, ie, the 5 which have the highest $|\widehat{\text{corr}}(v, q_{jL}(z)|$. Such modification was also used in [30].

A.1.2. *Penalty level choice for single outcome.* In every case when a single outcome variable is considered in isolation (this includes the reduced form selection step and the selection step corresponding to Φ_{K1}), Lasso is implemented with penalty λ as described in [8]. For ease of reference, note that [8] suggest λ given by $2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1-\alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} > 1, \alpha_{\text{Lasso}} \to 0$ are tuning parameters. In every instance in this paper, $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05$ are used.

A.1.3. *Penalty level choice for $\Phi_{K,\text{Simple}}$.* In this case, K Lasso regressions are run simultaneously. In this case, for all $\varphi \in \Phi_K$, λ is given by $2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05/K$ are used.

A.1.4. *Penalty level choice and implementation for $\Phi_{K,\text{Span}}$.* When the Span option is used, $\Phi_{K,Span}$ is decomposed $\Phi_{K,\text{ Span}} = \Phi_{K1} \cup \Phi_{K2} \cup \Phi_{K3}$. Each component has a corresponding penalty level applied to all φ within that component. On the first component, $\lambda_{\Phi_{K1}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05$. On the second component, $\lambda_{\Phi_{K2}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1-\alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05/K$. On the third component, $\lambda_{\Phi_{K3}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05/K$.

The following procedure is used for approximating I_{Φ_K} in the case that a component of Φ_K contains a continuum of test functions. For each $j \leqslant L$, a Lasso regression $\check{\varphi}_j \in \Phi_{K3}$ which is more likely to select $q_{jL(z)}$ than other $\varphi \in \Phi_K$. Specifically, for each j, $\check{\varphi}_j$ is set to the linear combination of $p_{1K}, ..., p_{KK}$ with highest marginal correlation to q_{jL}. Then the approximation to the first stage model selection step proceeds by using $\check{I}_{\Phi_{K3}} = \bigcup_{j \leqslant L} I_{\check{\varphi}_j(x)}$ in place of $I_{\Phi_{K3}}$.

A.1.5. *Penalty level choice for $\Phi_{K,\text{Span-Conservative}}$.* When the Conservative Span option is used, $\Phi_{K,\text{ Span-Conservative}}$ is decomposed $\Phi_{K,\text{ Span-Conservative}} = \Phi_{K1} \cup \Phi_{K2} \cup \Phi_{K3}$. Each component again has a corresponding penalty level applied to all φ within that component. On the first component, $\lambda_{\Phi_{K1}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05$. On the second component, $\lambda_{\Phi_{K2}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01$ and $\alpha_{\text{Lasso}} = .05/K$. On the third component, $\lambda_{\Phi_{K3}} = 2c_{\text{Lasso}}F^{-1}_{\text{N}(0,1)}(1 - \alpha_{\text{Lasso}}/L)$ where $c_{\text{Lasso}} = 1.01K^{1/2}$ and $\alpha_{\text{Lasso}} = .05$.

In order to approximate the variables selected on the continuum of Lasso estimates indexed by Φ_{K3}, the identical procedure with the Span option above is used. Note that the only difference between the Conservative Span option and the Span option is in $\lambda_{\Phi_{K3}}$.

A.2. p^K **Implementation Details.**

In every simulation and in the empirical example, p^K is constructed using a cubic B-spline expansion. For fixed K, the approximating dictionary is chosen according to the following procedure. Knots points $t_1, ..., t_{K-3}$ are chosen according to the following rule. Set

$$t_{\max} = \text{quantile}_{0.95}(|x_1|, ..., |x_n|) \text{ and } t_{\min} = -t_{\max}.$$

Let $\Delta_k = t_k - t_{k-1}$. For constants $c_1, c_2 \geqslant 0$ set

$$\Delta_k = c_1 + c_2|(K-2)/2 - k|$$

for $k = 2, ..., K-3$.

The constants c_1, c_2 serve to insert more knot points where the density of x is higher. The choices for c_1, c_2 are determined uniquely by the condition that $c_1 = 2c_2$ and that the endpoints satisfy $t_1 = t_{\min}$ and $t_{K-3} = t_{\max}$. Next, the B-spline formulation used here is given by the recursive formulation. Set

$$B_{k,0}(x) = \mathbf{1}_{t_k \leqslant x < t_{k+1}}.$$

Set $B_{k,0} = 0$ for k outside of $1, ..., K-3$. In addition, for spline order $o > 0$,

$$B_{k,o}(x) = \frac{x - t_k}{t_{k+o} - t_k} B_{k,o-1} + \frac{t_{k+o+1} - x}{t_{k+o+1} - t_{k+1}} B_{k+1,o-1}.$$

Set $(p_{1,K}(x), ..., p_{K-3,K}(x)) = (B_{1,3}(x), ...B_{K-3,3}(x))$. The dictionary is completed by adding the additional terms $p_{K-2,K}(x) = x,\ p_{K-1,K}(x) = x^2,\ p_{K,K}(x) = x^3$.

$\widehat{K}$ is chosen according to the following procedure. First, an initial set of terms $q^{\text{initial}}(z) \subseteq q^L(z)$ is selected. In each case, $q^{\text{initial}}(z)$ contains the terms I_{RF}. That is, the terms selected in a Lasso regression y on $q^L(z)$. Next, an initial value $\widehat{K}_0 \leqslant 2\lfloor n^{1/3} \rfloor$ is chosen to minimize BIC using $(p^K(x), q^{\text{initial}}(z))$. In the simulation study, $\widehat{K}_0$ is constrained to be $\geqslant 5$. Finally, in order to ensure undersmoothing, $\widehat{K}$ is set to $\widehat{K} = \lfloor (\log_{10}(n))\widehat{K}_0 \rfloor$.

A.3. **Targeted Undersmoothing Implementation Details.** The following procedure is used to estimate the Targeted Undersmoothing (TU; specifically TU(1); see [30]) confidence intervals for θ_0. For each $I \subseteq \{1, ..., p\}$ let $\widehat{\text{CI}}_{K,I}(\theta_0)$ be the corresponding confidence interval for θ_0 using K terms and the components of q^L corresponding to I. Then the full TU confidence interval is defined by the convex hull of $\cup_{j \leqslant p} \widehat{\text{CI}}_{\widehat{K}, I_{\text{RF}} \cup \{j\}}(\theta_0)$. In this implementation, a truncated TU confidence interval is calculated instead: $\cup_{j \leqslant 2s_0} \widehat{\text{CI}}_{\widehat{K}, I_{\text{RF}} \cup \{j\}}(\theta_0)$. This is done be the simulation run time reduces to the order of a day (from the order of a month), and therefore helps facilitate easier replicability. Changing the code to calculate the full TU confidence intervals is trivial. This also highlights that computing speed is another advantage of the Post-Nonparametric Double procedure relative to TU in certain settings. In terms of approximation error, the full TU estimator was implemented for the case $n = 100$, $p = 50$ for 1000 replications. The full TU confidence intervals as well as the truncated TU confidence intervals each made 9 false rejections. In addition, the average interval length for the full TU intervals was 1.740 while the average interval length for the truncated TU intervals was 1.722. Therefore, the truncated and full TU confidence intervals show very similar performance in this instance.

Appendix B. Proofs

B.1. **Preliminary Setup and Additional Notation.** Throughout the course of the proof, as much reference as possible is made to results in [45], [15]. This is done in order to maximize clarity and to present a better picture of the overall argument. In many cases, appealing directly to arguments in [45] is possible because many of the bounds required for deriving asymptotic normality for series estimators depend only on properties of $\widehat{g}$, g_0, p^K and D. Less direct appeal to bounds in the original Post-Double Selection argument is possible, since those arguments do not track K, and do not have notions of quantities stemming from Φ_K like α_ρ, α_Φ. However, the main idea of decomposing p^K into components in the span of, and orthogonal to q^L, remains as a theme throughout the proofs.

For any function φ, let $\varphi(X)$ denote the vector $[\varphi(x_1), \varphi(x_2), ..., \varphi(x_n)]'$. Similarly, let $\phi_{q^L}\varphi(Z) = [\pi_{q^L}\varphi(z_1), \pi_{q^L}\varphi(z_2), ..., \pi_{q^L}\varphi(z_n)]'$. In addition, define the following quantities.

1. Let m be the $n \times K$ matrix $m = \pi_{q^L} p^K(Z) = [\pi_{q^L} p_{1K}(Z), ..., \pi_{q^L} p_{KK}(Z)]$
3. Let $W = P - m$
4. Let $\widehat{\Omega} = n^{-1} P' \tilde{M} P$
5. Let $\Omega = n^{-1}\mathrm{E}[W'W]$
6. Let $\bar{\Omega} = n^{-1} W'W$
7. Let m be partitioned $m = [m_1, ..., m_K]$
8. Let W be partitioned $m = [W_1, ..., W_K]$
9. For any φ, let $R_\varphi = Q(\beta_{\varphi,L} - \beta_{\varphi,L,s_0})$
10. Let $R_y = Q(\beta_{y,L} - \beta_{y,L,s_0})$
11. For any φ, let $U_\varphi = \varphi(X) - Q\beta_{\varphi,L}$
12. Let $U_y = Y - Q\beta_{y,L}$
13. Let $F = V^{-1/2}$
14. Let $\varphi_a(x)$ be the function such that $\pi_{q^L}\varphi_a(Z) = FA'm$
15. Let $m_a = FA'm$
16. Let $W_a = \varphi_a(X) - m_a$.

Assume without loss of generality that $B_K = \mathrm{Id}_K$, the identity matrix of order K. The reason this is without loss of generality is that dictionary p^K is used only in the post-selection estimation, while Φ_K is used for first stage model selection. In addition, assume without loss of generality that $\Omega = \mathrm{Id}_K$.

Throughout the exposition, there is a common naming convention for various regression coefficients. Quantities of the form $\widehat{\beta}_{v,I}$ always denotes the sample regression coefficients from regressing the variable v on the components specified by I. This implies that the quantities $\widehat{\beta}_{\varphi,I_{\varphi,L}} = \widehat{\beta}_{\varphi,L,\text{Post-Lasso}}$ are equivalent, since the specified components being regressed on are the same. In addition, $\widehat{\beta}_{\varphi,I_{\Phi_K+\mathrm{RF}}} = \widehat{\beta}_{\varphi,\tilde{q}} = \widehat{\beta}_{\varphi(X),I_{\Phi_K+\mathrm{RF}}}$ are equivalent. Next, quantities of the form $\beta_{v,L}$ and β_{v,L,s_0} without a hat accent are population quantities and are defined in the text above.

B.2. **Preliminary Lemmas.**

Lemma 1. *Under the assumptions of Theorem 1,*

1. $J_1 := \max_{k \leqslant K} n^{-1/2} \|Q' W_k\|_\infty = O_p(\big(\log(KL)^{1/2}\big))$
2. $J_2 := n^{-1/2} \|Q' \mathcal{E}\|_\infty = O_p(\big(\log(L)^{1/2}\big))$
3. $J_3 := n^{-1/2} \|R_m' \mathcal{E}\|_2 = O_p((K K^{\alpha_\rho} L^{-\alpha_z}))$
4. $J_4 := n^{-1/2} \|R_{h_0}' W\|_2 = O_p(\big(n^{1/2} \zeta_0(K) L^{-\alpha_z}\big))$
5. $J_5 := \max_{k \leqslant K} n^{-1/2} \|\mathcal{M} m_k\|_2$
   $= O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0^{1/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
6. $J_6 := n^{-1/2} \|\mathcal{M} h_0(Z)\|_2 = O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0^{1/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
7. $J_7 := \max_{k \leqslant K} \|\widehat{\beta}_{m_k, I_{\Phi_K + RF}} - \beta_{p_{kK}, L, s_0}\|_1$
   $= O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0 K^{\alpha_{I_\Phi}/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
8. $J_8 := \|\widehat{\beta}_{h_0, I_{\Phi_K + RF}} - \beta_{h_0, L, s_0}\|_1$
   $= O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0 K^{\alpha_{I_\Phi}/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
9. $J_9 := \max_{k \leqslant K} \|\widehat{\beta}_{W_k, I_{\Phi_K + RF}}\|_1 = O_p(\Big(n^{-1/2} s_0^{1/2} K^{\alpha_{I_\Phi}/2} \log(KL)^{1/2}\Big))$
10. $J_{10} := \|\widehat{\beta}_{\mathcal{E}, I_{\Phi_K + RF}}\|_1 = O_p(\Big(n^{-1/2} s_0^{1/2} K^{\alpha_{I_\Phi}/2} \log(L)^{1/2}\Big))$
11. $J_{11} := n^{-1/2} \|Q' W_a\|_\infty = O_p(\big(\log(KL)^{1/2}\big))$
12. $J_{12} := n^{-1/2} \|R_{m_a}' \mathcal{E}\|_2 = O_p((K K^{\alpha_\rho} L^{-\alpha_z}))$
13. $J_{13} := n^{-1/2} \|\mathcal{M} m_a\|_2 = O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0^{1/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
14. $J_{14} := \|\widehat{\beta}_{m_a, I_{\Phi_K + RF}} - \beta_{\varphi_a, L, s_0}\|_1$
   $= O_p(\Big(n^{-1/2} K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0 K^{\alpha_{I_\Phi}/2} \log(L)^{1/2} + L^{-\alpha_z} K^{\alpha_\rho}\Big))$
15. $J_{15} := \|\widehat{\beta}_{W_a, I_{\Phi_K + RF}}\|_1 = O_p(\Big(n^{-1/2} s_0^{1/2} K^{\alpha_{I_\Phi}/2} \log(KL)^{1/2}\Big))$
15. $J_{16} := n^{-1} \|R_m' W\|_{\mathcal{F}} = O_p(K^{1/2} \zeta_0(K) K^{\alpha_\rho} L^{-\alpha_z})$.

Proof.

Statement 1. By Lemma 5 of [15], two conditions which together are sufficient for $\max_{k \leqslant K j \leqslant L} \frac{|Q_j' W_k|}{\sqrt{\sum_{i=1}^n q_{jL}(z_i)^2 W_{ki}^2}} = O_p((\log KL)^{1/2})$ are that $\max_{k \leqslant K, j \leqslant L} \frac{\mathrm{E}[|q_{jL}(z)|^3 |W_{ik}|^3]^{1/3}}{\mathrm{E}[q_{jL}(z)^2 W_{ik}^2]^{1/2}} = O(\zeta_0(K))$ and the rate condition $\log KL = o(\zeta_0(K)^{-1} n^{1/3})$. Note that $\mathrm{E}[q_{jL}(z)^2 W_{ik}^2]^{1/2}$ is bounded away from zero by assumption. In addition, by Hölder's inequality, $\mathrm{E}[|q_{jL}(z)|^3 |W_{ik}|^3] \leqslant \mathrm{E}[|q_{jL}(z)|^3] \zeta_0(K)^3$.

This implies that the first condition holds. The second condition is given in the assumptions.

Statement 2. Follows similarly as Statement 1.

Statement 3. This statement follows directly from the fact that $\mathrm{E}[\varepsilon|x,z]=0$, $\mathrm{E}[\varepsilon^2|x,z]$ bounded, along with $\dim(R_m'\mathcal{E})=K$ and $\|R_{m_k}\|_\infty = O(L^{-\alpha_z})$, allowing the use of the K-dimensional Chebyshev Inequality.

Statement 4. $\|R_{h_0}'W\|_2 = \|\sum_i R_{h_0,i}W_i\|_2 \leqslant O(L^{-\alpha_z})\zeta_0(K)$ by the facts that $R_{h_0}=O(L^{-\alpha_z})$ and $\|W_i\|_2 \leqslant \zeta_0(K)$.

Statement 5.

First note that the following two hold.

1. For any $\varphi \in \Phi_K$, $\mathcal{M}\pi_{q^L}\varphi(Z) = \mathcal{M}R_\varphi + \mathcal{M}(Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\varphi,L}})$.
2. For any $g \in \mathrm{LinSpan}(p^K)$, and any corresponding expansion $g=\eta_1\varphi_1 + ... + \eta_{k_g}\varphi_{k_g} + r_g$ with $\eta_1,...\eta_{k_g}\in\mathbb{R}, \varphi_1,...\varphi_{k_g}\in\Phi_K$,

$$\|\mathcal{M}\pi_{q^L}g(Z)\|_2 \leqslant \|\eta\|_1 \max_{\varphi\in\{\varphi_1,...,\varphi_{k_g}\}} \|Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\varphi,L}}\|_2 + \|\eta\|_1\|R_g\|_2 + \|r_g(Z)\|_2.$$

To show the first of the above two statements, for each $\varphi\in\Phi_K$, note that

$$\begin{aligned}
\mathcal{M}\pi_{q^L}\varphi(Z) &= \mathcal{M}\mathcal{M}\pi_{q^L}\varphi(Z)\\
&= \mathcal{M}(\pi_{q^L}\varphi(Z) - \mathcal{P}\pi_{q^L}\varphi(Z))\\
&= \mathcal{M}(Q\beta_{\varphi,L} - \mathcal{P}(\varphi(X)-U_\varphi))\\
&= \mathcal{M}(Q\beta_{\varphi,L} - Q\widehat{\beta}_{\varphi,I_{\Phi_K+\mathrm{RF}}} + \mathcal{P}U_\varphi)\\
&= \mathcal{M}R_\varphi + \mathcal{M}(Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\Phi_K+\mathrm{RF}}}) + \mathcal{M}\mathcal{P}U_\varphi\\
&= \mathcal{M}R_\varphi + \mathcal{M}(Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\Phi_K+\mathrm{RF}}})\\
&= \mathcal{M}R_\varphi + \mathcal{M}(Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\varphi,L}}) + \underbrace{\mathcal{M}(Q\widehat{\beta}_{\varphi,I_{\varphi,L}} - Q\widehat{\beta}_{\varphi,I_{\Phi_K+\mathrm{RF}}})}\\
&\qquad\qquad = \mathcal{M}(\mathcal{P}_{I_{\varphi,L}}\varphi(X) - \mathcal{P}\varphi(X))\\
&\qquad\qquad = \mathcal{M}\mathcal{P}(\mathcal{P}_{I_{\varphi,L}}\varphi(X) - \varphi(X))\\
&\qquad\qquad = 0\\
\Rightarrow \mathcal{M}\pi_{q^L}\varphi(Z) &= \mathcal{M}R_\varphi + \mathcal{M}(Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\varphi,L}}).
\end{aligned}$$

This establishes the first claim. Now turn to the second claim. Note that using the density assumption, there are $\varphi_1,...,\varphi_{k_g}$ and a vector $\eta=(\eta_1,...,\eta_{k_g})$ such that $g=\eta_1\varphi_1+...+\eta_{k_g}\varphi_{k_g}+r_g$ for some remainder r_g, sufficiently small. Then

$$\|\mathcal{M}\pi_{q^L}g(Z)\|_2 = \|\eta_1\mathcal{M}\pi_{q^L}\varphi_1(X) + ... + \eta_{k_g}\mathcal{M}\pi_{q^L}\varphi_{k_g}(Z) + \mathcal{M}\pi_{q^L}r_g(Z)\|_2$$

Next, looking at each φ in the above expansion (ie each $\varphi\in\{\varphi_1,...,\varphi_{k_g}\}$) and combining the above expression gives

$$\|\mathcal{M}\pi_{q^L}g(Z)\|_2 = \|\eta_1\mathcal{M}R_{\varphi_1} + \eta_1\mathcal{M}(Q\beta_{\varphi_1,L,s_0} - Q\widehat{\beta}_{\varphi_1,I_{\varphi_1}}) + ...$$
$$... + \eta_{k_g}\mathcal{M}R_{\varphi_{k_g}} + \eta_{k_g}\mathcal{M}(Q\beta_{\varphi_{k_g},L,s_0} - Q\widehat{\beta}_{\varphi_{k_g},I_{\varphi_{k_g}}}) + \mathcal{M}\pi_{q^L}r_g(Z)\|_2.$$

Applying Hölder's inequality and the fact that $\mathcal{M}$ is a projection (and hence non-expansive) gives the bound

$$\leqslant \|\eta\|_1 \max_{\varphi\in\{\varphi_1,...,\varphi_{k_g}\}} \|Q\beta_{\varphi,L,s_0} - Q\widehat{\beta}_{\varphi,I_{\varphi,L}}\|_2 + \|\eta\|_1\|R_g\|_2 + \|r_g(Z)\|_2.$$

These can then be applied directly to $n^{-1/2}\|\mathcal{M}m_k\|_2$. Under Assumption 9, note that for m_k, the corresponding η and R_{m_k} satisfy $\|\eta\|_1 \leqslant O(K^{\alpha_\rho})$ and $\|R_{m_k}\|_2 \leqslant n^{1/2}O(L^{-\alpha_z}K^{\alpha_\rho})$. Then we have the bound

$$\|\mathcal{M}\pi_{q^L}g(Z)\|_2 = O_p(K^{\alpha_\rho}K^{\alpha_\Phi/2}s_0^{1/2}\log(L)^{1/2} + n^{1/2}L^{-\alpha_z}K^{\alpha_\rho}).$$

Under Assumption 10, note that for each m_k, taking $\eta = 1$ and $R_{m_k} = 0$ are feasible by assumption. The result follows.

Statement 6.

$$\begin{aligned}
n^{-1/2}\|\mathcal{M}h_0(Z)\|_2 &= n^{-1/2}\|\mathcal{M}(Q\beta_{h_0,L,s_0} + R_{h_0})\|_2) \\
&\leqslant n^{-1/2}(\|\mathcal{M}Q\beta_{h_0,L,s_0}\|_2 + \|\mathcal{M}R_{h_0}\|_2) \\
&\leqslant n^{-1/2}(\|\mathcal{M}Q(\beta_{g_0,L,s_0} + \beta_{h_0,L,s_0} - Q\beta_{g_0,L,s_0})\|_2 + \|\mathcal{M}R_{h_0})\|_2 \\
&= n^{-1/2}(\|\mathcal{M}(Q\beta_{y,L,s_0} - Q\beta_{g_0,L,s_0})\|_2 + \|\mathcal{M}R_{h_0}\|_2) \\
&\leqslant n^{-1/2}(\|\mathcal{M}Q\beta_{y,L,s_0}\|_2 + \|\mathcal{M}Q\beta_{g_0,L,s_0}\|_2 + \|\mathcal{M}R_{h_0}\|_2) \\
&= n^{-1/2}(\|\mathcal{M}\pi_{q^L}y(Z)\|_2 + \|\mathcal{M}\pi_{q^L}g_0(Z)\|_2 + \|\mathcal{M}R_{h_0}\|_2)
\end{aligned}$$

The first two terms above, $n^{-1/2}(\|\mathcal{M}Q\pi_{q^L}y(Z)\|_2 + n^{-1/2}\|\mathcal{M}\pi_{q^L}g_0(Z)\|_2)$, are $O_p(K^{\alpha_\rho}K^{\alpha_\Phi/2}n^{-1/2}s_0^{1/2}\log(L)^{1/2} + L^{-\alpha_z}K^{\alpha_\rho})$ by the same reasoning as Statement 6. In addition $n^{-1/2}\|\mathcal{M}R_{h_0}\|_2 \leqslant n^{-1/2}\|R_{h_0}\|_2 = O(L^{-\alpha_z})$ by assumption. This gives

$$n^{-1/2}\|\mathcal{M}h_0(Z)\|_2 = O_p(K^{\alpha_\rho}K^{\alpha_\Phi/2}n^{-1/2}s_0^{1/2}\log(L)^{1/2} + L^{-\alpha_z}K^{\alpha_\rho}).$$

Statement 7.

$$
\begin{aligned}
&\|\widehat{\beta}_{m_k, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0}\|_1 \\
&\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2} \|\widehat{\beta}_{m_k, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0}\|_2 \\
&= |I_{\Phi_K+\mathrm{RF}}|^{1/2} \left((\widehat{\beta}_{m_k, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0})' (\widehat{\beta}_{m_k, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0}) \right)^{1/2} \\
&\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2} O_p(1) \left((\widehat{\beta}_{m_k, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0})' (Q'_{I_{\Phi_K+\mathrm{RF}}} Q_{I_{\Phi_K+\mathrm{RF}}}/n) (\widehat{\beta}_{g, I_{\Phi_K+\mathrm{RF}}} - \beta_{p_{kK},L,s_0}) \right)^{1/2} \\
&= |I_{\Phi_K+\mathrm{RF}}|^{1/2} O_p(1) n^{-1/2} \|\mathcal{P} m_k - Q \beta_{p_{kK},L,s_0}\|_2 \\
&= |I_{\Phi_K+\mathrm{RF}}|^{1/2} O_p(1) n^{-1/2} \|m_k - \mathcal{M} m_k - Q \beta_{p_{kK},L,s_0}\|_2 \\
&= |I_{\Phi_K+\mathrm{RF}}|^{1/2} O_p(1) n^{-1/2} \| - \mathcal{M} m_k + R_{m_k}\|_2 \\
&\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2} O_p(1) (J_5 + O(L^{-\alpha_z})) \\
&= O_p(s_0^{1/2} K^{\alpha_{I_\Phi}/2}) (J_5 + O(L^{-\alpha_z})) \\
&= O_p(s_0^{\alpha_{I_\Phi}/2+1/2}) O_p(K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0^{1/2} \log(L)^{1/2} + n^{1/2} L^{-\alpha_z} K^{\alpha_\rho}) \\
&= O_p(K^{\alpha_\rho} K^{\alpha_\Phi/2} s_0^1 K^{\alpha_{I_\Phi}/2} \log(L)^{1/2} + n^{1/2} L^{-\alpha_z} K^{\alpha_\rho}).
\end{aligned}
$$

Statement 8. Proven analogously to Statement 7.

Statement 9.

$$
\begin{aligned}
&\max_{k \leqslant K} \|\widehat{\beta}_{W_k, I_{\Phi_K+\mathrm{RF}}}\|_1 \\
&= \max_{k \leqslant K} \|(\tilde{Q}'\tilde{Q})^{-1} \tilde{Q}' W_k\|_1 \\
&\leqslant ||I_{\Phi_K+\mathrm{RF}}|^{1/2} \max_{k \leqslant K} \|(\tilde{Q}'\tilde{Q})^{-1} \tilde{Q}' W_k\|_2 \\
&\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2} \kappa_{\min}^{-1/2}(I_{\Phi_K+\mathrm{RF}}|) \max_{k \leqslant K} \|n^{-1} Q' W_k\|_\infty \\
&= O_p(s_0^{1/2} K^{\alpha_{I_\Phi}/2} \cdot 1 \cdot n^{-1/2} \log(KL)^{1/2}).
\end{aligned}
$$

Statement 10. Proven analogously to Statement 9.

Statements 11-15. Proven analogously to Statements 1,3,5,7,9.

Statement 16.

$$
\begin{aligned}
n^{-1} \Big\| \sum_{i=1}^n W_i' R_{m,i} \Big\|_{\mathcal{F}} &= n^{-1} \left(\sum_k \|W' R_{m_k}\|_2^2 \right)^{1/2} \\
&\leqslant n^{-1} \left(\sum_k n^2 \zeta_0(K)^2 \|R_{m_k}^2\|_\infty \right)^{1/2}.
\end{aligned}
$$

By the density assumption, $\|R_{m_k}\|_\infty \leqslant K^{\alpha_\rho} L^{-\alpha_z}$. This then implies that

$$n^{-1}\Big\| \sum_{i=1}^{n} W_i' R_{m,i} \Big\|_{\mathcal{F}} \leqslant K^{1/2} K^{\alpha_\rho} L^{-\alpha_z}.$$

□

Lemma 2.

1. $\Xi_1 := n^{-1}\|W'\mathcal{P}W\|_{\mathcal{F}} \leqslant n^{-1/2} K J_9 J_1$
2. $\Xi_2 := n^{-1}\|m'\mathcal{M}m\|_{\mathcal{F}} \leqslant K J_5^2$
3. $\Xi_3 := n^{-1}\|m'\mathcal{M}W\|_{\mathcal{F}} \leqslant J_{16} + n^{-1/2} K J_7 J_1$
4. $\Xi_4 := n^{-1/2}\|m'\mathcal{M}h_0(Z)\|_2 \leqslant n^{1/2} K^{1/2} J_5 J_6$
5. $\Xi_5 := n^{-1/2}\|W'\mathcal{M}h_0(Z)\|_2 \leqslant J_4 + K^{1/2} J_8 J_1$
6. $\Xi_6 := n^{-1/2}\|W'\mathcal{P}\mathcal{E}\|_2 \leqslant K^{1/2} J_9 J_2$
7. $\Xi_7 := n^{-1/2}\|m'\mathcal{M}\mathcal{E}\|_2 \leqslant J_4 + K^{1/2} J_7 J_2$
8. $\Xi_8 := n^{-1/2}|m_a'\mathcal{M}h_0(Z)| \leqslant n^{1/2} J_5 J_{13}$
9. $\Xi_9 := n^{-1/2}|W_a'\mathcal{M}h_0(Z)| \leqslant J_{12} + J_{14} J_1$
10. $\Xi_{10} := n^{-1/2}|W_a'\mathcal{P}\mathcal{E}| \leqslant J_9 J_{11}$
11. $\Xi_{11} := n^{-1/2}|m_a'\mathcal{M}\mathcal{E}| \leqslant J_{12} + J_7 J_{11}$.

Proof.

Statement 1.

$$\begin{aligned}
\left(n^{-1}\|W'\mathcal{P}W\|_{\mathcal{F}}\right)^2 &= \sum_{k,\bar{k}\leqslant K} (n^{-1} W_k'\mathcal{P}W_{\bar{k}})^2 = \\
&= \sum_{k,\bar{k}\leqslant K} (n^{-1}\widehat{\beta}'_{W_k, I_{\Phi_K + RF}} Q' W_{\bar{k}})^2 \\
&\leqslant \sum_{k,\bar{k}\leqslant K} \|n^{-1/2}\widehat{\beta}_{W_k, I_{\Phi_K + RF}}\|_1^2 \|n^{-1/2} Q' W_{\bar{k}}\|_\infty^2 \\
&= \left(\sum_{k\leqslant K} \|n^{-1/2}\widehat{\beta}_{W_k, I_{\Phi_K + RF}}\|_1^2\right)\left(\sum_{\bar{k}\leqslant K} \|n^{-1/2} Q' W_{\bar{k}}\|_\infty^2\right) \\
&\leqslant K \cdot n^{-1} J_9^2 \cdot K \cdot J_1^2 \\
\\
&\Rightarrow n^{-1}\|W'\mathcal{P}W\|_{\mathcal{F}} \leqslant n^{-1/2} K J_1 J_9.
\end{aligned}$$

Statement 2.

$$\begin{aligned}\left(n^{-1}\|m'\mathcal{M}m\|_{\mathcal{F}}\right)^2 &= \sum_{k,\bar{k}\leqslant K}(n^{-1}m_k'\mathcal{M}m_{\bar{k}})^2 \leqslant \sum_{k,\bar{k}\leqslant K}\|n^{-1/2}\mathcal{M}m_k\|_2^2\|n^{-1/2}\mathcal{M}m_{\bar{k}}\|_2^2 \\ &= \left(\sum_{k\leqslant K}\|n^{-1/2}\mathcal{M}m_k\|_2^2\right)^2 \leqslant K^2J_5^4 \\ \Rightarrow n^{-1}\|m'\mathcal{M}m\|_{\mathcal{F}} &\leqslant KJ_5^2.\end{aligned}$$

Statement 3.

$$\begin{aligned}n^{-1}\|m'\mathcal{M}W\|_{\mathcal{F}} &= n^{-1}\|m'W/n - m'\mathcal{P}W\|_{\mathcal{F}} \\ &= n^{-1}\|R_m'W + (Q\beta_{p^K,L,s_0})'W - m'\mathcal{P}W\|_{\mathcal{F}} \\ &= n^{-1}\|R_m'W + (Q\beta_{p^K,L,s_0})'W - (Q\widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'W\|_{\mathcal{F}} \\ &= n^{-1}\|R_m'W + (\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'W\|_{\mathcal{F}} \\ &\leqslant n^{-1}\|R_m'W\|_{\mathcal{F}} + n^{-1}\|(\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'W\|_{\mathcal{F}}.\end{aligned}$$

Then the first term in the last line is bounded above as $n^{-1}\|R_m'W\|_{\mathcal{F}} = J_{16}$ while the second term has

$$\begin{aligned}&\left(n^{-1}\|(\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'W\|_{\mathcal{F}}\right)^2 \\ &= n^{-2}\sum_{k,\bar{k}\leqslant K}((\beta_{p_k,L,s_0} - \widehat{\beta}_{m_k,I_{\Phi_K+\mathrm{RF}}})'Q'W_{\bar{k}})^2 \\ &\leqslant n^{-2}\sum_{k,\bar{k}\leqslant K}\|\beta_{p_k,L,s_0} - \widehat{\beta}_{m_k,I_{\Phi_K+\mathrm{RF}}}\|_1^2\|Q'W_{\bar{k}}\|_\infty^2 \\ &= n^{-1}\left(\sum_{k\leqslant K}\|\beta_{p_k,L,s_0} - \widehat{\beta}_{m_k,I_{\Phi_K+\mathrm{RF}}}\|_1^2\right)\left(\sum_{\bar{k}\leqslant K}\|n^{-1/2}Q'W_{\bar{k}}\|_\infty^2\right) \\ &\leqslant n^{-1}K\cdot J_7^2\cdot K\cdot J_1^2.\end{aligned}$$

Therefore, $n^{-1}\|m'\mathcal{M}W\|_{\mathcal{F}} \leqslant J_{16} + n^{-1/2}KJ_7J_1$.

Statement 4.

$$\begin{aligned}n^{-1/2}\|m'\mathcal{M}h_0(Z)\|_2 &\leqslant n^{1/2}\|n^{-1/2}\mathcal{M}h_0(Z)\|_2K^{1/2}\max_{k\leqslant K}n^{-1/2}\|m'\mathcal{M}\|_2 \\ &\leqslant n^{1/2}K^{1/2}J_5J_6.\end{aligned}$$

Statement 5.

$$\begin{aligned}
n^{-1/2}\|W'\mathcal{M}h_0(Z)\|_2 &= n^{-1/2}\|W'h_0(Z) - W'\mathcal{P}h_0(Z)\|_2 \\
&= n^{-1/2}\|W'h_0(Z) - W'Q\widehat{\beta}_{h_0(Z),I_{\Phi_K+\mathrm{RF}}}\|_2 \\
&= n^{-1/2}\|W'R_{h_0} + W'Q\beta_{h_0,L,s_0} - W'Q\widehat{\beta}_{h_0(Z),I_{\Phi_K+\mathrm{RF}}}\|_2 \\
&= n^{-1/2}\left(\|W'h_0(Z)\|_2 + \|(\widehat{\beta}_{h_0(Z),I_{\Phi_K+\mathrm{RF}}} - \beta)'Q'W\|_2\right) \\
&\leqslant J_4 + K^{1/2}\max_{k\leqslant K}\|\widehat{\beta}_{h_0(Z),I_{\Phi_K+\mathrm{RF}}} - \beta_{h_0,L,s_0}\|_1\|n^{-1/2}Q'W_k\|_\infty \\
&\leqslant J_4 + K^{1/2}J_8J_1.
\end{aligned}$$

Statement 6.

$$\begin{aligned}
\left(n^{-1/2}\|W'\mathcal{P}W\|_2\right)^2 &= n^{-1}\sum_{k\leqslant K}(W_k'\mathcal{P}\mathcal{E})^2 \\
&= n^{-1}\sum_{k\leqslant K}(\widehat{\beta}'_{W_k,I_{\Phi_K+RF}}Q'\mathcal{E})^2 \\
&\leqslant \sum_{k\leqslant K}\|\widehat{\beta}_{W_k,I_{\Phi_K+RF}}\|_1^2\|n^{-1/2}Q'\mathcal{E}\|_\infty^2 \\
&\leqslant K\cdot J_9^2\cdot J_2^2 \\
&\Rightarrow n^{-1/2}\|W'\mathcal{P}\mathcal{E}\|_{\mathcal{F}} \leqslant K^{1/2}J_9J_2.
\end{aligned}$$

Statement 7.

$$\begin{aligned}
n^{-1/2}\|m'\mathcal{M}\mathcal{E}\|_2 &= n^{-1/2}\|m'\mathcal{E}/n - m'\mathcal{P}\mathcal{E}\|_{\mathcal{F}} \\
&= n^{-1/2}\|R_m'\mathcal{E} + (Q\beta_{p^K,L,s_0})'W - m'\mathcal{P}\mathcal{E}\|_{\mathcal{F}} \\
&= n^{-1/2}\|R_m'\mathcal{E} + (Q\beta_{p^K,L,s_0})'W - (Q\widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'\mathcal{E}\|_2 \\
&= n^{-1/2}\|R_m'\mathcal{E} + (\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'\mathcal{E}\|_2 \\
&\leqslant n^{-1/2}\|R_m'\mathcal{E}\|_2 + n^{-1/2}\|(\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'\mathcal{E}\|_2.
\end{aligned}$$

Then the first term in the last line is bounded above as $n^{-1/2}\|R_m'\mathcal{E}\|_2 = J_4$. Turning to the second term,

$$\begin{aligned}
&\left(n^{-1/2}\|(\beta_{p^K,L,s_0} - \widehat{\beta}_{m,I_{\Phi_K+\mathrm{RF}}})'Q'\mathcal{E}\|_2\right)^2 \\
&= n^{-1}\sum_{k\leqslant K}((\beta_{p_k,L,s_0} - \widehat{\beta}_{m_k,I_{\Phi_K+\mathrm{RF}}})'Q'\mathcal{E})^2 \\
&\leqslant \sum_{k\leqslant K}\|\beta_{p_k,L,s_0} - \widehat{\beta}_{m_k,I_{\Phi_K+\mathrm{RF}}}\|_1^2\|n^{-1/2}Q'\mathcal{E}\|_\infty^2 \\
&\leqslant K\cdot J_7^2\cdot J_2^2.
\end{aligned}$$

Therefore, $n^{-1/2}\|m'\mathcal{M}\mathcal{E}\|_2 \leqslant J_4 + K^{1/2}J_7J_2$.

Statements 8-11.

The argument is identical to the argument for Statements 4-7, adjusting appropriately for the fact that m_a is 1-dimensional rather than K-dimensional. □

The following corollaries follow directly from assumed rate conditions and the above bounds. These are used in the proof of Theorems 1 and 2.

Corollary 1. *Under the assumptions of Theorem 1,*

1. $\Xi_1 + \Xi_2 + \Xi_3 = O_p(n^{-1/2}\zeta_0(K)K^{1/2})$
2. $n^{-1/2}(\Xi_4 + \Xi_5 + \Xi_6 + \Xi_7) = O_p(n^{-1/2}K^{1/2} + K^{-\alpha_{g_0}})$.

Corollary 2. *Under the assumptions of Theorem 2,*

1. $n^{-1/2}\zeta_0(K)K^{1/2}(\Xi_4 + \Xi_5 + \Xi_6 + \Xi_7) = o_p(1)$.
2. $\Xi_8 + \Xi_9 + \Xi_{10} + \Xi_{11} = o_p(1)$.

B.3. **Proof of Theorem 1.**

Lemma 3.

1. $\|\widehat{\Omega} - \Omega\|_{\mathcal{F}} \leqslant O_p(\zeta_0(K)K^{1/2}n^{-1/2}) + \Xi_1 + \Xi_2 + \Xi_3 = o_p(1)$
2. $\|\widehat{\Omega}^{-1} - \Omega^{-1}\|_{2\to 2} = O_p(\zeta_0(K)K^{1/2}n^{-1/2}) + O(\Xi_1 + \Xi_2 + \Xi_3)$.

Proof. The argument in Theorem 1 of [45] gives the bound $\|\bar{\Omega} - \Omega\|_{\mathcal{F}} = O_{\mathrm{P}}(\zeta_0(K)K^{1/2}n^{-1/2})$. Next, using the decomposition, $P = m + W$, write $\widehat{\Omega} = (m+W)'\mathcal{M}(m+W)/n = W'W/n - W'(\mathrm{Id}_n - \mathcal{M})W/n + m'\mathcal{M}m/n + 2m'\mathcal{M}W/n$. By triangle inquality, $\|\bar{\Omega} - \widehat{\Omega}\|_{\mathcal{F}} \leqslant \|W'\mathcal{P}W/n\|_{\mathcal{F}} + \|m'\mathcal{M}m/n\|_{\mathcal{F}} + \|2m'\mathcal{M}W/n\|_{\mathcal{F}} = \Xi_1 + \Xi_2 + \Xi_3$. Bounds for each of the three above terms are established above along with the assumed rate conditions give $\|\bar{\Omega} - \widehat{\Omega}\|_{\mathcal{F}} = o_p(1)$. The last statement holds by applying an expansion of the matrix inversion function around Id_K.

$$\widehat{\Omega}^{-1} = (\mathrm{Id}_K - (\mathrm{Id}_K - \widehat{\Omega}))^{-1} = \mathrm{Id}_K + (\mathrm{Id}_K - \widehat{\Omega}) + (\mathrm{Id}_K - \widehat{\Omega})^2 + ...$$

The sum given above is with probability $\to 1$ absolutely convergent relative to the Frobenius norm $\mathcal{F}$. In addition, by the bound $\|\cdot\|_{2\to 2} \leqslant \|\cdot\|_{\mathcal{F}}$, we have $\|\widehat{\Omega}^{-1} - \mathrm{Id}_K\|_{2\to 2} \leqslant \|\widehat{\Omega} - \mathrm{Id}_K\|_{\mathcal{F}} \leqslant \|\mathrm{Id}_K - \widehat{\Omega}\|_{\mathcal{F}} + \|\mathrm{Id}_K - \widehat{\Omega}\|_{\mathcal{F}}^2 + ... = O_p(\zeta_0(K)K^{1/2}n^{-1/2}) + O(\Xi_1 + \Xi_2 + \Xi_3)$.

□

Note that since Ω has minimal eigenvalues bounded from below by assumption, it follows that $\widehat{\Omega}$ and $\bar{\Omega}$ are invertible with probability approaching 1. The reference [45] works on the event $1_n := \{\lambda_{\min}(\widehat{\Omega}) > 1/2\}$ and later uses the fact that this event has probability $\to 1$. This fact is used several times, however its use is only implicitly in reference to arguments in [45].

Lemma 4. $\|\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}\mathcal{E}\|_2 = O_p(n^{-1/2}K^{1/2})$.

Proof.

$$\begin{aligned}\|\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}\mathcal{E}\|_2 &\leqslant \|\widehat{\Omega}^{-1}\|_{2\to 2}n^{-1}\|P'\mathcal{M}\mathcal{E}\|_2 \\ &\leqslant \|\widehat{\Omega}^{-1}\|_{2\to 2}(n^{-1}\|W'\mathcal{E}\|_2 + n^{-1/2}\Xi_6 + n^{-1/2}\Xi_7)\end{aligned}$$

$\|W'\mathcal{E}\|_2 = O_p(n^{-1/2}K^{1/2} = O_p(n^{-1/2})$ by arguments in [45]. Bounds for $n^{-1/2}\Xi_6 + n^{-1/2}\Xi_7$ follows from the previous Lemmas and from the assumed rate conditions. □

Lemma 5. $\|\widehat{\Omega}^{-1}P'\mathcal{M}(g_0(X) - P\beta_{g_0,K})/n\|_2 = O_p(K^{-\alpha_{g_0}})$.

Proof.

$$\begin{aligned}\|\widehat{\Omega}^{-1}P'\mathcal{M}(g_0(X) - P\beta_{g_0,K})/n\|_2 &= [(g_0(X) - P\beta)'\mathcal{M}P\widehat{\Omega}^{-1}P'\mathcal{M}(g_0(X) - P\beta)/n]^{1/2} \\ &= O_p(1)[(g_0(X) - P\beta_{g_0,K})'(g_0(Z) - P\beta_{g_0,K})/n]^{1/2} \\ &= O_p(K^{-\alpha_{g_0}})\end{aligned}$$

by assumption on $(g_0(X) - P\beta_{g_0,K})$ and idempotency of $\mathcal{M}P\widehat{\Omega}^{-1}P'\mathcal{M} = \mathcal{M}P(P'\mathcal{M}\mathcal{M}P)^{-1}\mathcal{M}$. □

Lemma 6. $\|\widehat{\Omega}^{-1}P'\mathcal{M}h_0(Z)/n\|_2 = o_p(n^{-1/2})$.

Proof. $\widehat{\Omega}$ has eigenvalues bounded below and above with probability approaching 1. Then,

$$\begin{aligned}\|\widehat{\Omega}^{-1}P'\mathcal{M}h_0(Z)/n\|_2 &\leqslant O_p(1)\|P'\mathcal{M}h_0(Z)/n\|_2 \\ &= O_p(1)\|(m+W)'\mathcal{M}h_0(Z)/n\|_2 \\ &= n^{-1/2}O_p(1)n^{-1/2}\|(m+W)'\mathcal{M}h_0(Z)/n\|_2 \\ &\leqslant n^{-1/2}O_p(1)(n^{-1/2}\|m'\mathcal{M}h_0(Z)\|_2 + n^{-1/2}\|W'\mathcal{M}h_0(Z)\|_2) \\ &= n^{-1/2}O_p(1)(\Xi_4 + \Xi_5) \\ &= n^{-1/2}O_p(1)o_p(1).\end{aligned}$$

□

Lemma 7. $\|\widehat{\beta}_g - \beta_{g_0,K}\|_2 = O_p(n^{-1/2}K^{1/2} + K^{-\alpha_{g_0}})$.

Proof. Note that $([\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_g - \beta_{g_0,K}) = n^{-1}\widehat{\Omega}^{-1}P'\mathcal{M}\mathcal{E} + n^{-1}\widehat{\Omega}^{-1}P'M_{\widehat{I}}(g_0(X) - P\beta_{g_0,K}) + n^{-1}\widehat{\Omega}^{-1}P'\mathcal{M}h_0(Z)$. Triangle inequality in conjuction with the bounds described in the previous three lemmas give the result. □

The final statement of Theorem 1 follows from the bound on $\|\widehat{\beta}_g - \beta_{g_0,K}\|_2$ using the arguments in [45]. ■

B.4. **Proof of Theorem 2.** Recall that $F = V^{-1/2}$. Let $\bar{g} = p^K(x)'\beta_{g_0,K}$ and decompose the quantity $n^{1/2}F[a(\widehat{g}) - a(g_0)]$ by

$$\begin{aligned}n^{1/2}F[a(\widehat{g}) - a(g_0)] = n^{1/2}F[&a(\widehat{g}) - a(g_0) + D(\widehat{g}) - D(g_0) \\ &+ D(\bar{g}) - D(\widehat{g}) \\ &+ D(g_0) - D(\bar{g})].\end{aligned}$$

Lemma 8. $n^{1/2}F[D(\bar{g}) - D(g_0)] = O(n^{1/2}K^{-\alpha_{g_0}})$.

Proof. This follows from arguments given in the proof of Theorem 2 in [45]. Note that the statement does not contain any reference to random quantities. □

Lemma 9. $|n^{1/2}F[a(\widehat{g}) - a(g) - D(\widehat{g}) + D(g)| = o_p(1)$.

Proof. Bounds on $|\widehat{g} - g|_d$ given by Theorem 1 imply that $|n^{1/2}F[a(\widehat{g}) - a(g_0) - D(\widehat{g}) + D(g_0)| \leqslant Cn^{1/2}|\widehat{g} - g_0|_d^2 = O_p(n^{1/2}(n^{-1/2}\zeta_d(K)K^{1/2} + K^{-\alpha_{g_0}})^2) = o_p(1)$. This is again identical to the reasoning given in Theorem 2 in [45], since that references uses only a bound on $|\widehat{g} - g|_d$ to prove the analogous result. □

The last step is to show that $n^{1/2}F[D(\widehat{g}) - D(\bar{g})] \to_d \mathrm{N}(0,1)$.

Lemma 10. $n^{1/2}F[D(\widehat{g}) - D(\bar{g})] \to_d \mathrm{N}(0,1)$.

Proof. Note that $D(\widehat{g})$ can be expanded

$$\begin{aligned}
D(\widehat{g}) &= D(p^K(x)'[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_g) = D(p^K(x)'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}Y) \\
&= D(p^K(x)'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}(g_0(X) + h_0(Z) + \mathcal{E})) \\
&= D(p^K(x))'\widehat{\Omega}^{-1}n^{-1}(g_0(X) + h_0(Z) + \mathcal{E}) \\
&= A'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}(g_0(X) + h_0(Z) + \mathcal{E}) \\
&= A'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}g_0(X) + A'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}h_0(Z) + A'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}\mathcal{E}.
\end{aligned}$$

In addition, $D(\bar{g}) = D(p^K(x)'\beta_{g_0,K}) = A'\beta_{g_0,K}$ gives

$$\begin{aligned}
n^{1/2}F[D(\widehat{g}) - D(\bar{g})] &= n^{1/2}FA'[\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}g_0(X) - \beta_{g_0,K}] \\
&\quad + n^{1/2}FA'[\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}(h_0(Z) + \mathcal{E})].
\end{aligned}$$

The above equation gives a decomposition of the right hand side into two terms, which are next bounded separately. Before proceeding, note that the following bounds $\|FA\|_2 = O(1)$, $\|FA'\widehat{\Omega}^{-1}\|_2 = O_p(1)$, $\|FA'\widehat{\Omega}^{-1/2}\|_2 = O_p(1)$, $\|FA'\Omega^{-1}\|_2 = O_p(1)$, $\|FA'\Omega^{-1/2}\|_2 = O(1)$ all hold by arguments in [45]. Consider the first term.

$$\begin{aligned}
&|n^{1/2}FA'[n^{-1}\widehat{\Omega}^{-1}P'\mathcal{M}g_0(X) - \beta_{g_0,K}]| \\
&= |\sqrt{n}FA'[(P'\mathcal{M}P/n)^{-1}P'\mathcal{M}(G - P\beta)/n]| \\
&\leqslant \|FA'\widehat{\Omega}^{-1}P'\mathcal{M}/\sqrt{n}\|_2\|g_0(X) - P\beta_{g_0,K}\|_2 \\
&\leqslant \|FA'\widehat{\Omega}^{-1}P'\mathcal{M}/\sqrt{n}\|_2\sqrt{n}\max_{i\leqslant n}|g(x_i) - \bar{g}(x_i)| \\
&= \|FA'\widehat{\Omega}^{-1/2}\|_2\sqrt{n}\max_{i\leqslant n}|g(x_i) - \bar{g}(x_i)| \\
&\leqslant \|FA'\widehat{\Omega}^{-1/2}\|_2\sqrt{n}|g - \bar{g}|_0 \\
&= O_p(1)O_p(\sqrt{n}K^{-\alpha}) \\
&= o_p(1).
\end{aligned}$$

Next, consider $n^{1/2}FA'\widehat{\Omega}^{-1}n^{-1}P'\mathcal{M}(h_0(Z)+\mathcal{E})$. To handle this term, first bound

$$\begin{aligned}
&|n^{-1/2}FA'(\widehat{\Omega}^{-1} - \Omega^{-1})P'\mathcal{M}(h_0(Z) + \mathcal{E})| \\
&\leqslant n^{-1/2}\|FA'(\widehat{\Omega}^{-1} - \Omega^{-1})\|_2\|P'\mathcal{M}(h_0(Z) + \mathcal{E})\|_2 \\
&= \|FA'(\widehat{\Omega}^{-1} - \Omega^{-1})\|_2(n^{-1/2}\|P'\mathcal{M}(h_0(Z) + \mathcal{E})\|_2) \\
&= \|FA'(\widehat{\Omega}^{-1} - \Omega^{-1})\|_2(n^{-1/2}\|(m + W)'\mathcal{M}(h_0(Z) + \mathcal{E})\|_2) \\
&\leqslant \|FA'(\widehat{\Omega}^{-1} - \Omega^{-1})\|_2(\Xi_4 + \Xi_5 + \Xi_6 + \Xi_7) \\
&\leqslant \|\widehat{\Omega}^{-1} - \Omega^{-1}\|_{2\to 2}\|FA'\|_2(\Xi_4 + \Xi_5 + \Xi_6 + \Xi_7) \\
&= o_p(1).
\end{aligned}$$

Next consider the last remaining term for which a central limit result will be shown.

$$\begin{aligned}
&\sqrt{n}FA'\Omega^{-1}P'\mathcal{M}(h_0(Z)+\mathcal{E})/n \\
&= \sqrt{n}FA'\Omega^{-1}(W+m)'\mathcal{M}(h_0(Z)+\mathcal{E})/n \\
&= \sqrt{n}(FA'\Omega^{-1}W+m_a)'\mathcal{M}(h_0(Z)+\mathcal{E})/n \\
&= \sqrt{n}FA'\Omega^{-1}W\mathcal{M}\mathcal{E}+\sqrt{n}m_a'\mathcal{M}(h_0(Z)+\mathcal{E})/n+\sqrt{n}W_a'\mathcal{M}h_0(Z)/n \\
&= \sqrt{n}FA'\Omega^{-1}W\mathcal{E}-\sqrt{n}FA'\Omega^{-1}W\mathcal{P}\mathcal{E}+\sqrt{n}m_a'\mathcal{M}(h_0(Z)+\mathcal{E})/n+\sqrt{n}W_a'\mathcal{M}h_0(Z)/n \\
&= \sqrt{n}FA'\Omega^{-1}W'\mathcal{E}/n+o_p(1).
\end{aligned}$$

Note that the last $o_p(1)$ bound in the equation array above holds by the fact that $|\sqrt{n}FA'\Omega^{-1}W\mathcal{P}\mathcal{E}+\sqrt{n}m_a'\mathcal{M}(h_0(Z)+\mathcal{E})/n+\sqrt{n}W_a'\mathcal{M}h_0(Z)/n| \leqslant \Xi_8+\Xi_9+\Xi_{10}+\Xi_{11}$. The term $\sqrt{n}FA'\Omega^{-1}W'\mathcal{E}/n$ satisfies the conditions Lindbergh-Feller Central Limit Theorem, by arguments given in [45]. □

The previous three lemmas prove that $n^{1/2}F[a(\widehat{g})-a(g_0)] \to \mathrm{N}(0,1)$.

The next set of arguments bound $\widehat{V}-V$. For ν as in the statement of Assumption 14, Define the event $\mathcal{A}_g = \{|\widehat{g}-g_0|_d < \nu/2\}$. Define $\widehat{u} = 1_{\mathcal{A}_g}\widehat{\Omega}^{-1}\widehat{A}F$ and $u = 1_{\mathcal{A}_g}\Omega^{-1}AF$. In addition, define $\bar{\Sigma} = \sum_i W_iW_i'\varepsilon_i^2/n$, an infeasible sample analogue of Σ.

Lemma 11.

1. $\|\widehat{A}-A\|_2 = o_p(1)$
2. $\|\widehat{u}-u\|_2 = o_p(1)$
3. $\|\bar{\Sigma}-\Sigma\|_{\mathcal{F}} = o_p(1)$
4. $|\widehat{u}\bar{\Sigma}\widehat{u}-\widehat{u}'\Sigma\widehat{u}| = o_p(1)$.

Proof. **Statement 1.** In the case that $a(g)$ is linear in g, then $a(p'\beta) = A'\beta \implies \widehat{A} = A$. Therefore, consider the case that $a(g)$ is not linear in g. Using arguments identical to those in [45], $1_{\mathcal{A}_g} = 1$ with probability $\to 1$, and

$$1_{\mathcal{A}_g}\|\widehat{A}-A\|_2 \leqslant C\cdot\zeta_d(K)|\widehat{g}-g|_d.$$

Statement 2. This follows from arguments in [45].

Statement 3. This follows from arguments in [45].

Statement 4. An immediate implication of Statement 3 is that $1_{\mathcal{A}_g}|\widehat{u}\bar{\Sigma}\widehat{u}-\widehat{u}'\Sigma\widehat{u}| = |\widehat{u}'(\bar{\Sigma}-\Sigma)\widehat{u}| \leqslant \|\widehat{u}\|_2^2\|\bar{\Sigma}-\Sigma\|_{2\to 2}^2 = O_p(1)o_p(1)$.

□

Lemma 12. $\max_{i\leqslant n}|h_0(z_i)-\widehat{h}(z_i)| = o_p(1)$.

Proof. First note that

$$\begin{aligned}
\max_i|h_0(z_i)-\widehat{h}(z_i)| &\leqslant \max_i|h_0(z_i)-q^L(z_i)'\beta_{h_0,L,s_0}| \\
&\quad+\max_i|q^L(z_i)'[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_h-q^L(z_i)'\beta_{h_0,L,s_0}|.
\end{aligned}$$

The first term has the bound $\max_i |h_0(z_i) - q(x_i)'\eta| = O_p(L^{-\alpha_z})$ by assumption. Next,

$$\begin{aligned}\max_i |q^L(z_i)'[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_h - q^L(z_i)'\beta_{h_0,L,s_0}| &= \max_i |q^L(z_i)'([\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_h - \beta_{h_0,L,s_0})| \\ &\leqslant \max_i \|q^L(z_i)\|_\infty \|[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_h - \beta_{h_0,L,s_0}\|_1\end{aligned}$$

Then,

$$\begin{aligned}\|[\widehat{\beta}_{y,(\tilde{p},\tilde{q})}]_h - \beta_{h_0,L,s_0}\|_1 &= \|\widehat{\beta}_{y-\widehat{g},I_{\Phi_K+\mathrm{RF}}} - \beta_{h_0,L,s_0}\|_1 \\ &= \|\widehat{\beta}_{g_0,I_{\Phi_K+\mathrm{RF}}} + \widehat{\beta}_{h_0,I_{\Phi_K+\mathrm{RF}}} + \widehat{\beta}_{\varepsilon,I_{\Phi_K+\mathrm{RF}}} - \widehat{\beta}_{\widehat{g},I_{\Phi_K+\mathrm{RF}}} - \beta_{h_0,L,s_0}\|_1. \\ &\leqslant \|\widehat{\beta}_{h_0,I_{\Phi_K+\mathrm{RF}}} - \beta_{h_0,L,s_0}\|_1 + \|\widehat{\beta}_{\varepsilon,I_{\Phi_K+\mathrm{RF}}}\|_1 + \|\widehat{\beta}_{g_0-\widehat{g},I_{\Phi_K+\mathrm{RF}}}\|_1 \\ &= J_8 + J_{10} + \|\widehat{\beta}_{g_0-\widehat{g},I_{\Phi_K+\mathrm{RF}}}\|_1.\end{aligned}$$

Next,

$$\begin{aligned}&\|\widehat{\beta}_{g_0-\widehat{g},I_{\Phi_K+\mathrm{RF}}}\|_1 \\ &\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2}\|\widehat{\beta}_{g_0-\widehat{g},I_{\Phi_K+\mathrm{RF}}}\|_2 \\ &= |I_{\Phi_K+\mathrm{RF}}|^{1/2}\|(Q'_{I_{\Phi_K+\mathrm{RF}}}Q_{I_{\Phi_K+\mathrm{RF}}}/n)^{-1}Q'_{I_{\Phi_K+\mathrm{RF}}}(g_0(X) - \widehat{g}(X))/n\|_2 \\ &\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2}\kappa_{\min}(|I_{\Phi_K+\mathrm{RF}}|)^{-1}\|Q'_{I_{\Phi_K+\mathrm{RF}}}(g_0(X) - \widehat{g}(X))/n\|_2 \\ &\leqslant |I_{\Phi_K+\mathrm{RF}}|^{1/2}\kappa_{\min}(|I_{\Phi_K+\mathrm{RF}}|)^{-1}|I_{\Phi_K+\mathrm{RF}}|^{1/2}\|Q'_{I_{\Phi_K+\mathrm{RF}}}(g_0(X) - \widehat{g}(X))/n\|_\infty \\ &\leqslant |I_{\Phi_K+\mathrm{RF}}|\kappa_{\min}(|I_{\Phi_K+\mathrm{RF}}|)^{-1}\|Q'_{I_{\Phi_K+\mathrm{RF}}}(g_0(X) - \widehat{g}(X))/n\|_\infty \\ &= |I_{\Phi_K+\mathrm{RF}}|\kappa_{\min}(|I_{\Phi_K+\mathrm{RF}}|)^{-1}\left(\max_j n^{-1}\sum_{i=1}^n |q_{jL}(z_i)|\right)\|g_0(X) - \widehat{g}(X)\|_\infty \\ &= O_p(s_0^1 K^{\alpha_{I_\Phi}})O_p(1)o_p(n^{-1/2}\zeta_0(K)K^{1/2} + K^{-\alpha_{g_0}}).\end{aligned}$$

Putting these together, it follows from the assumed rate conditions that

$$\max_i |h_0(z_i) - \widehat{h}(z_i)| = o_p(1).$$

□

Next, let $\Delta_{g_0 i} = g_0(x_i) - \widehat{g}(x_i)$ and $\Delta_{h_0 i} = h_0(z_i) - \widehat{h}(z_i)$. Then above lemma states $\max_{i\leqslant n}\Delta_{h_0 i} = o_p(1)$. In addition $\max_{i\leqslant n}|\Delta_{g_0 i}| \leqslant |\widehat{g} - g|_0 = o_p(1)$. Let $\omega_i^2 = u'W_iW_i'u$ and $\widehat{\omega}_i^2 = \widehat{u}'W_iW_i'\widehat{u}$.

Lemma 13. $|F\widehat{V}F - \widehat{u}\bar{\Sigma}\widehat{u}| = o_p(1)$.

Proof.

$$\begin{aligned}1_{\mathcal{A}_g}|F\widehat{V}F - \widehat{u}'\bar{\Sigma}\widehat{u}| = |\widehat{u}'(\widehat{\Sigma} - \bar{\Sigma})\widehat{u}| &= \left|\sum_{i=1}^n \widehat{u}'\widehat{W}_i\widehat{W}_i'\widehat{\varepsilon}_i^2\widehat{u}/n - \sum_{i=1}^n \widehat{u}'W_iW_i'\varepsilon_i^2\widehat{u}/n\right| \\ &\leqslant \left|\sum_{i=1}^n \omega_i^2(\widehat{\varepsilon}_i^2 - \varepsilon_i^2)/n\right| + \left|\sum_{i=1}^n(\widehat{\omega}_i^2 - \omega_i^2)\widehat{\varepsilon}_i^2/n\right|.\end{aligned}$$

Both terms on the right hand side will be bounded. Consider the first term. Expanding $(\widehat{\varepsilon}_i^2 - \varepsilon_i^2)$ gives

$$\left|\sum_{i=1}^n \omega_i^2(\widehat{\varepsilon}_i^2 - \varepsilon_i^2)/n\right| \leqslant \left|\sum_{i=1}^n \omega_i^2\Delta_{1i}^2/n\right| + \left|\sum_{i=1}^n \omega_i^2\Delta_{2i}^2/n\right| + \left|\sum_{i=1}^n \omega_i^2\Delta_{1i}\Delta 2i/n\right| + 2\left|\sum_{i=1}^n \omega_i^2\Delta_{1i}\varepsilon_i/n\right| + 2\left|\sum_{i=1}^n \omega_i^2\Delta_{2i}\varepsilon_i/n\right|.$$

Note that $\sum_{i=1}^n \omega_i^2/n, \sum_{i=1}^n \omega_i^2|\varepsilon_i| = O_p(1)$ by arguments is in [45]. The five terms above are then bounded in order of their appearence by

$$\sum_{i=1}^n \omega_i^2\Delta_{1i}^2/n \leqslant \max_{i\leqslant n}|\Delta_{1i}| \sum_{i=1}^n \omega_i^2/n = o_p(1)O_p(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{2i}^2/n \leqslant \max_{i\leqslant n}|\Delta_{2i}| \sum_{i=1}^n \omega_i^2|\varepsilon_i|/n = o_p(1)O_p(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{1i}\Delta_{2i}/n \leqslant \max_{i\leqslant n}|\Delta_{1i}| \max_{i\leqslant n}|\Delta_{2i}| \sum_{i=1}^n \omega_i^2/n = o_p(1)O_p(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{1i}\varepsilon_i/n \leqslant \max_{i\leqslant n}|\Delta_{1i}| \sum_{i=1}^n \omega_i^2|\varepsilon_i|/n = o_p(1)O_p(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{2i}\varepsilon_i/n \leqslant \max_{i\leqslant n}|\Delta_{2i}| \sum_{i=1}^n \omega_i^2|\varepsilon_i|/n = o_p(1)O_p(1).$$

The second term is bounded by

$$\left|\sum_{i=1}^n \widehat{u}(\widehat{W}_i\widehat{W}_i' - W_iW_i')\widehat{\varepsilon}_i^2\widehat{u}/n\right| \leqslant \max_{i\leqslant n}|\varepsilon_i^2| \left|\sum_{i=1}^n \widehat{u}(\widehat{W}_i\widehat{W}_i' - W_iW_i')\widehat{u}/n\right|$$
$$\leqslant \max_{i\leqslant n}|\widehat{\varepsilon}_i^2|\|\widehat{u}\|_2^2\| \sum_{i=1}^n (\widehat{W}_i\widehat{W}_i' - W_iW_i')/n\|_{2\to 2} = \max_{i\leqslant n}|\widehat{\varepsilon}_i^2|\|\widehat{u}\|_2^2\|\widehat{\Omega} - \bar{\Omega}\|_{2\to 2}$$
$$\leqslant \left(\max_{i\leqslant n}|\varepsilon_i^2| + \max_{i\leqslant n}|\widehat{\varepsilon}_i^2 - \varepsilon_i^2|\right)\|\widehat{u}\|_2^2\|\widehat{\Omega} - \bar{\Omega}\|_{2\to 2}$$
$$= \Big(O_p(n^{2/\delta}) + o_p(1)\Big)O_p(1)(O_p(\zeta_0(K)n^{-1/2}K^{1/2}) + \Xi_1 + \Xi_2 + \Xi_3)$$
$$= o_p(1).$$

where the last bounds come from the rate condition in Assumption 9 and $\max_{i\leqslant n}|\widehat{\varepsilon}_i^2 - \varepsilon_i^2| = o_p(1)$ by $\max_{i\leqslant n}|\Delta_{1i}| + |\Delta_{2i}| = o_p(1)$. □

These results give the conclusion that

$$n^{1/2}\widehat{V}^{-1/2}(\widehat{\theta} - \theta) = n^{1/2}(F\widehat{V}F)^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0,1).$$

Calculations which give the rates of convergence in each of the cases of Assumption 17 or of Assumption 18, as well as the proof of the second statement of Theorem 2, use the same arguments as in [45]. This concludes the proof. ■

References

[1] P. Aghion and P.W. Howitt. *The Economics of Growth*. MIT Press, 2008.

[2] Donald D.W. Andrews and Yoon-Jae Whang. Additive interactive regression models: Circumvention of the curse of dimensionality. *Econometric Theory*, 6(4):466479, 12 1990.

[3] Donald W. K. Andrews. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–345, 1991.

[4] J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.

[5] J. Bai and S. Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24, 2009.

[6] R.J. Barro and J.W. Lee. Data set for a panel of 139 countries. NBER,http://www.nber.org/pub/barro.lee.html (1994).

[7] Robert J. Barro and Jong-Wha Lee. Losers and winners in economic growth. Working Paper 4341, National Bureau of Economic Research, April 1993.

[8] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429, 2012. Arxiv, 2010.

[9] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ArXiv, 2009.

[10] A. Belloni, V. Chernozhukov, I. Fernndez-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[11] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. 2010 arXiv:[math.ST], http://arxiv.org/abs/1012.1297, 2010.

[12] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2013.

[13] Alexandre Belloni and Victor Chernozhukov. *High Dimensional Sparse Econometric Models: An Introduction*, pages 121–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[14] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345 – 366, 2015. High Dimensional Problems in Econometrics.

[15] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls with an application to abortion on crime. *Review of Economic Studies*, 81(2):608–650, 2014.

[16] Alexandre Belloni, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605, 2016.

[17] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[18] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[19] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *Ann. Statist.*, 17(2):453–510, 06 1989.

[20] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[21] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.

[22] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[23] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[24] Been-Lon Chen. Economic growth : Robert J. Barro and Xavier Sala-i-Martin, (McGraw-Hill, 1995), 539 pp. *Journal of Economic Dynamics and Control*, 21(4-5):895–898, May 1997.

[25] R. Chen, W. Härdle, O. B. Linton, and E. Severance-Lossin. Nonparametric estimation of additive separable regression models. In Wolfgang Härdle and Michael G. Schimek, editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 247–265, Heidelberg, 1996. Physica-Verlag HD.

[26] Norbert Christopeit and Stefan G. N. Hoderlein. Local partitioned regression. *Econometrica*, 74(3):787–817, 2006.
[27] Dennis D. Cox. Approximation of least squares regression on nested subspaces. *Ann. Statist.*, 16(2):713–732, 06 1988.
[28] Brian J. Eastwood and A. Ronald Gallant. Adaptive rules for seminonparametric estimators that achieve asymptotic normality. *Econometric Theory*, 7(3):307–340, 1991.
[29] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
[30] C. Hansen, D. Kozbur, and S. Misra. Targeted Undersmoothing. *ArXiv e-prints*, June 2017.
[31] Trevor Hastie and Robert Tibshirani. [generalized additive models]: Rejoinder. *Statist. Sci.*, 1(3):314–318, 08 1986.
[32] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
[33] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.
[34] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 08 2010.
[35] Guido Imbens and Keisuke Hirano. The propensity score with continuous treatments. 2004.
[36] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.
[37] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.
[38] Keith Knight. Shrinkage estimation for nearly singular designs. *Econometric Theory*, 24:323–337, 2008.
[39] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.
[40] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.
[41] Qi Li and Jeffrey Scott Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press: Princeton, NJ, 2006.
[42] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
[43] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.
[44] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
[45] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.
[46] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.
[47] Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.
[48] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, June 2013.
[49] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.
[50] Eric Severance-Lossin and Stefan Sperlich. Estimation of derivatives for additive separable models. *Statistics*, 33(3):241–265, 1999.
[51] Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.
[52] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
[53] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
[54] Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014.

[55] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, May 2009.

[56] Lijian Yang, Stefan Sperlich, and Wolfgang Hrdle. Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference*, 115(2):521 – 542, 2003.

[57] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

[58] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[59] S. Zhou. Restricted eigenvalue conditions on subgaussian matrices. *ArXiv:0904.4723v2*, 2009.